云计算和大数据：

这一时代的两位王者

Big Data

大数据走向云计算

Cloud Computing

张德丰 | 编著

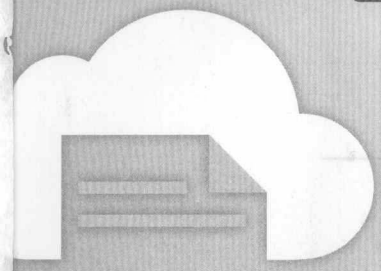
人民邮电出版社
POSTS & TELECOM PRESS

Big Data

大数据走向云计算

Cloud Computing

张德丰 编著



人民邮电出版社
北京

图书在版编目（CIP）数据

大数据走向云计算 / 张德丰编著. -- 北京 : 人民邮电出版社, 2014.4
ISBN 978-7-115-33986-7

Ⅰ. ①大… Ⅱ. ①张… Ⅲ. ①计算机网络－数据处理
Ⅳ. ①TP393

中国版本图书馆CIP数据核字(2013)第319222号

内 容 提 要

本书以 Hadoop 为铺垫，以概念、价值、动向及应用为导线，系统地介绍了大数据走向云计算的原理与技术。首先，介绍了云时代、大数据时代的基本内容，让读者了解到云计算、大数据各自的知识点；其次，介绍了大数据走向云端、云下的大数据工程，让读者领略到大数据在云的作用下的价值及应用；然后，介绍了搭建云计算开发环境、分布式文件系统、并行计算、分布式锁等内容，让读者认识到 Hadoop 组件的构建及使用；接着，介绍了数据挖掘、社会中的大数据等应用，让读者掌握到大数据走向云端的需求及益处。最后，总结介绍云下的大数据应用，让读者真正领会到大数据走向云端的实际效用。

本书适合于云计算、大数据初中级读者使用，也可作为大数据专业研究人员的参考资料。

◆ 编　　著　张德丰
责任编辑　刘　博
责任印制　彭志环　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：22.75　　2014 年 4 月第 1 版
字数：568 千字　　2014 年 4 月北京第 1 次印刷

定价：49.00 元

读者服务热线：(010) 81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前　言

随着互联网技术的不断发展而产生的云计算，已经成为国内外信息专家和学者们关注的热门话题。互联网就是“云”，云又由子云体系组成。未来的互联网可能是这样的一个情景，一些网站信息服务云，一些邮件服务云，一些企业成本管理云等。

子云体系的特征如下。

（1） 由大量具有相同需求目的的成员组成。

（2） 成员的数量可以随意增加或者减少。

（3） 计算与服务和成员管理是分开的，它们处于一个独立的源。成员从这个源里获得具体的计算与服务。

云系统按照服务目的可以分成两个子云系统：消费云和供应云。消费云是指使用计算能力的成员，成员包括消费单位以及他们所使用的软硬件设施等；供应云是指提供计算能力的成员，成员包括供应商以及他们提供的软硬件设施等。

而数据库是现在最领先的一个数据管理模式，它可以把数据进行很好的归类，进行非常快速的检索，过去 30 年我们生活在数据库的时代。

随着数据本身的改变，云所带来的副作用及云的使用者正在改变云的三大特征，即将产生一个新的数据云的时代，从过去数据库一家独大到新的数据云，会产生新的需求，产生更大、更快的数据，分布更广、更多样的数据，同时这些数据能够为千家万户，为所有的用户提供服务。当然，数据库并不会消失，数据库仍旧有它非常重要的作用。在很长的时间里，这两种技术会是共存的。还有两个非常重要的在业界的趋势，会帮助在更好的管理数据库的同时，能够迎接这个数据云时代的到来，而且使两边能够和谐的生存。

更重要的开源效果是对于数据云时代的帮助，数据管理云系统，现在走过一个分久必合到合久必分的转型的时代。在 30 年前可能是群雄混战，有很多的数据库产生，而在过去的 20 年、15 年，甲骨文逐渐一家做大，成为业界的领袖。但是现在这个时代，我们的“皇帝”也老了，新一代的技术产生，我们又进入一个群雄混战的时代。现在大家熟悉的大数据技术，包括 Hadoop 等，新的为开发者所欢迎的技术已经产生，而这样的技术大多数是以开源技术的形式出现的。开源就使得客户可以非常低的门槛应用到这个技术，不需要很多的初始投资，可以尝试这个新的技术到底是否满足自己的需求。百花齐放，开源就给百花齐放提供了一片土壤，看最后到底哪一朵花投其所好。

根据目前数据变化的趋势，传统的数据库已经 Hold 不住所有的应用了。数据变化趋势主要有：

第一是海量数据的需求。这些数据基本上是以每年成倍的速度进行发展，而对于大的数据量的分析需求往往也更细，而对它的门槛要求也更低，传统的数据库无法满足这种需求。

第二是对于快的需求。很多时候数据得到的同时，就希望有智能的产生，希望有反应，对应用希望能够直接产生效果。只能有低延迟，同时在数据流产生的时候就能够有 Action 的产生。

第三是过去的开发者决定着什么样的数据当先，而现在的开发者，包括移动、社交应用的开发者，往往需求是多样化的，而在这些多样化的需求里面，很多时候关系型数据库并不

是最优的解决方案。

正因为这些开发者的需求，使得各种各样的解决方案能够大行其道，包括一些大家非常熟悉的，都在互联网的应用当中，在移动和社交的应用当中有广泛的应用。随着大量、海量数据，以及实时快速、灵活的需求，同时客户也希望能够以自助型的形式得到应用，一个云的模式，让开发者能够自己部署他的数据系统，部署在混合云、虚拟架构之上。

在一个大数据文件系统中，往往对这些数据进行三种分析。

第一种，为需要进行实时分析。

第二种，为交互处理。

第三种，为批处理。

在这个大数据处理平台，需要有一个统一的基础架构，在这上面需要有三种不同的处理模式，满足平台的需要。Hadoop 在其中是非常重要的技术，但是并不是非常充分的技术。还需要很多新的技术，能够和 Hadoop 一起满足客户的需要。在这个大数据的平台之上，有一个数据分析的应用及数据展示的模式，可以提供给各种各样的用户。

根据大数据到云端的需求，我们编写了本书，本书主要内容包括：

第 1 章：介绍了云时代，主要包括云计算力量、云产生的背景、云计算简史、云研究趋势、云安全、云标准等内容。

第 2 章：介绍了大数据时代，主要包括大数据来源、大数据的价值、大数据技术、大数据变革、大数据转型、大数据应用等内容。

第 3 章：介绍了大数据走向云端，主要包括云计算与大数据的联系、数据向迁移云计算移到、云延伸、云计算与大数据挑战、机遇并存等内容。

第 4 章：介绍了云下的大数据工程，主要包括信息所需求的新生力量、信息系统工程、信息系统工程架构转变、商业变革因素、云计算机遇等内容。

第 5 章：介绍了搭建云计算开发环境，主要包括 Hadoop 环境搭建、HBase 环境搭建、ZooKeeper 环境搭建、Pig 环境搭建、MapReduce 概述等内容。

第 6 章：介绍了分布式文件系统，主要包括分布式文件系统概述、分布式文件系统类型、分布式结构数据表等内容。

第 7 章：介绍了并行计算，主要包括并行计算概述、MapReduce 基础、MapReduce 模板、MapReduce 计算流程等内容。

第 8 章：介绍了存储仓库，主要包括 HBase 基本特征、HBase 数据库、HBase 的模型及 HBase 的基本接口等内容。

第 9 章：介绍了分布式锁，主要包括 ZooKeeper 基本概述、Zookeeper 角色、ZooKeeper 接口与编程、ZooKeeper 的典型应用等内容。

第 10 章：介绍了数据挖掘，主要包括数据挖掘概述、PageRank 工具、关联分析、聚类分析、分类分析、异常挖掘、特异群组分析、矩估计等内容。

第 11 章：介绍了社会中的大数据，主要大数据在历史战争中应用、普适计算、数据应用于治国上、商务智能等内容。

第 12 章：介绍了云下的大数据应用，主要包括云计算如何实现价值、云计算与大数据的强强联合、大数据在证券中的应用及大数据在视频监控中应用等内容。

本书适用于云计算、大数据初、中、高级读者使用，也可作为研究大数据相关专业研究

人员的参考资料。

本书主要由张德丰编写，此外参加编写的还有刘泳、邓耀隆、李嘉乐、卢佳华、梁志成、高泳崇、曾虹雁、钟东山、邓秀乾、张金林、邓俊辉、李旭波、梁朗星和刘超。

由于时间仓促，加之编者水平有限，书中错误和疏漏之处在所难免。在此，诚恳地期望得到各领域的专家和广大读者的批评指正。

编　者

2013 年 11 月

目　录

第1章　云时代

大数据浪潮，汹涌来袭，与互联网的发明一样，这绝对不仅仅是信息技术领域的革命，更是在全球范围启动透明政府、加速企业创新、引领社会变革的利器。现代管理学之父德鲁克曾言，预测未来最好的方法，即是去创造未来。而“大数据战略”，则是当下领航全球的先机。

大数据，这一世界大潮的来龙去脉怎样呢？数据技术变革，何以能推动政府信息公开、透明和社会公正？何以促进行政管理和商业管理革新，并创造无限商机？又何以既便利又危及我们每个人的生活？Google、百度之类搜索服务，何以会不再有立足之地？

这是一个信息爆炸的时代，互联网上的信息正在以几何级数的速度增长。在这个大背景下，消耗CPU最多的计算逐渐从“提升软件本身性能”方面转移到信息处理方面。与此同时，摩尔定律似乎也不再像以前那么准确地发挥作用了。在这样的严峻形势下，各大厂商面临着极大的挑战——他们需要从TB及至PB级的数据中挖掘出有用的信息，并对这些海量的数据进行快捷、高效的处理。在这段特殊时期，Google公司以MapReduce为基石，结合CFS、Bigtable逐步发展成为全球互联网企业的领头羊。然而，出于技术保密的原因，Google公司并没有开源其MapReduce的实现细节，这使得人们无法深入了解和认识它。就在这时，一头神秘的大象——Hadoop从天而降，它的开源给人带来了新的希望、新的景象。

1.1　云计算力量

凭借着丰富的实践经验，IBM总结了企业进入云计算的四大路径。如果企业的目的是降低IT开支和复杂性，可能利用云计算，云化优化数据中心，这样就能满足降低IT成本和复杂性的要求。企业的目的是为了加速产品上市的速度，就可以利用云化的平台服务解决新服务快速上市的需求，从而增加利润和竞争优势。如果企业的目的是为了获得对企业软件和应用的及时访问，可以利用云技术获得对企业解决方案及时的访问，同时分析固定成本最小化。如果企业有创新的新业务模式，可成为云服务供应商。

IBM在这四种企业进入云计算的路径中都可以提供产品和服务。例如，在云服务供应商方面，IBM可以提供高级、可靠，安全、可扩展性的平台，使云服务的供应商可以很方便地创建管理和服务，为业务提供保障。

IBM在中国也实施了非常多的云计算案例，在过去的一两年中，帮助交通银行打造了银行核心业务的系统运维营，帮助七匹狼打造了商务平台，构建了中国第一个三维互联网孵化平台。在这里向大家介绍三个案例。

1．案例一

国内的某大型保险公司，在全国有30余家分公司，它采用两级应用架构，所有的分公司都有自己的IT应用设施和系统管理人员。这家企业希望利用数据中心整合，将大部分业务系统集中运行在统一的数据中心，统一的管理运维，实现资源共享，提高IT系统的资源利用率。IBM建立了云计算模式的数据中心，通过云计算工具的实施，实现资源的统一管理和调

度的自动化，对于客户来讲，通过建立共享的业务服务平台，为所有的公司提供业务测试服务，在降低业务成本的同时，提高了业务能力，利用虚拟化技术，实现了资源的利用率，实现分公司的运营。

2．案例二

关于公有云。在中国，很多人认为 IBM 公司主要是提供私有云的服务，IBM 公司也是全球最大的企业级的公有云的供应商，但是在中国，由于政策的限制，尚未提供公有云技术。这个案例就是有关公有云的案例，IBM 为一家全球最大的电信公司提供了公有云服务，提供的是 IBM 解决方案，对他的帮助是快速上市，如果他自己研发制造公有云平台，可能需要两年到三年的时间，而 IBM 给他提交，让他快速上市，比原有的方案提前了 12 个月到 18 个月，而且他可以立即享用 IBM 全球生态系统，同时也极大地降低了他们的风险，降低了他们的财务投入。

3．案例三

某大型化工制造业公司，它的工厂遍布亚洲和北美地区，其企业用于大量的支撑工厂运营的独立化系统，必须进行人工作业才能完成。系统一旦发生异常，工厂的运营效率就会大大降低。IBM 为这个化工企业量身打造了移动运维平台，将员工的资料、文件、档案，全部上传到云中心，简化了机制，高层人员随时随地就可以了解企业、管理企业。通过 IBM 实施云的方案，这个企业现有的 25 个独立工厂的自动化系统，整合到两个云中心，实现并强化了数据整合，资源优化，以及运营效率的提高，支持管理层通过设备了解工厂的运营状况。

1.2　云计算概述

云计算（cloud computing）是基于互联网的相关服务的增加、使用和交付模式，通常通过互联网来提供动态易扩展且经常是虚拟化的资源。

云计算是继 20 世纪 80 年代大型计算机到客户端-服务器的大转变之后的又一种巨变。用户不再需要了解“云”中基础设施的细节，不必具有相应的专业知识，也无需直接进行控制。云计算描述了一种基于互联网的新的 IT 服务增加、使用和交付模式。

云计算概况如图 1-1 所示。

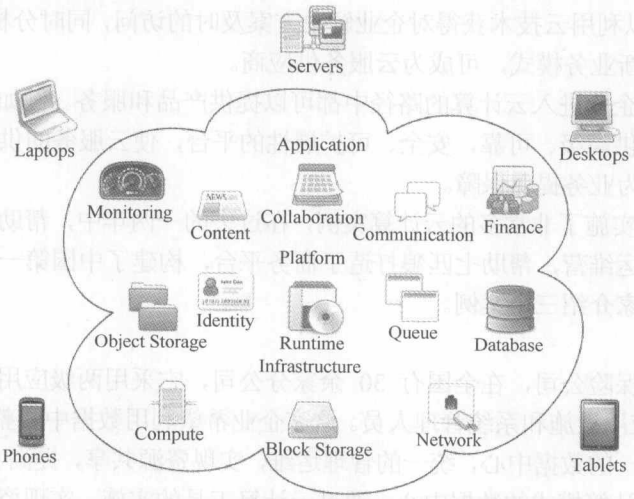


图 1-1　云计算概况图

用户通过浏览器、桌面应用程序或是移动应用程序来访问云的服务。推广者认为云计算使得企业能够更迅速地部署应用程序，并降低管理的复杂度和维护成本，以及允许 IT 资源的迅速重新分配以应对企业需求的快速改变。

云计算依赖资源的共享以达成规模经济，类似基础设施（如电力网）。服务提供者集成大量的资源供多个用户使用，用户可以轻易地请求（租借）更多资源，并随时调整使用量，将不需要的资源释放回整个架构，因此用户不需要因为偶尔大量的需求就购买大量的资源，仅需提升租借量，需求降低时便退租。服务提供者得以将目前无人租用的资源重新租给其他用户，甚至依照整体的需求量调整租金。

1.2.1　云基本特征

互联网上的云计算服务特征和自然界的云、水循环具有一定的相似性，因此，云是一个相当贴切的比喻。根据美国国家标准和技术研究院的定义，云计算服务应该具备如下几条特征：

① 随需自助服务；

② 随时随地用任何网络设备访问；

③ 多人共享资源池；

④ 快速重新布署灵活度；

⑤ 可被监控与量测的服务。

一般认为还有如下特征：

① 基于虚拟化技术快速部署资源或获得服务；

② 减少用户终端的处理负担；

③ 降低了用户对于 IT 专业知识的依赖。

核心特性主要有如下几点。

① 敏捷（Agility）。使用户得以快速且以低价格获得技术架构资源。

② 应用程序界面（API）的可达性。是指允许软件与云以类似“人机交互这种用户界面设施交互的方式”来交互。云计算系统典型的运用基于 REST（Representational State Transfer）网络架构的 API。

③ 在公有云中的传输模式中的支持已经转变为运营成本，故费用大幅下降。很显然的降低了进入门栏，这是由于典型的体系架构是由第三方提供，且无需一次性购买，且没有了罕见的集中计算任务的压力。称为计算资源包的通用计算基础上的原则在细粒度上基于用户的操作和更少的 IT 技能被内部实施。

④ 设备和本地依赖允许用户通过网页浏览器来获取资源，而无需关注用户自身是通过何种设备，或在何地介入资源（如 PC、移动设备等）。通常设施是在非本地的（典型的是由第三方提供的），并且通过因特网获取，用户可以从任何地方来连接。

- 一种称为多租户的软件架构技术允许在多用户池下共享资源与消耗。
- 体系结构的中央化使得本地的耗用更少（如不动产、电力等）。
- 峰值负载能力增加（用户无需建造最高可能的负载等级）。
- 原先利用率只有 10%～20%的系统利用效率增加了。

⑤ 如果使用多个冗余站点，则改进了可靠性，这允许我们设计云计算以符合商业一致性及灾备。

⑥ 可扩展性经由在合理粒度上按需的服务开通资源，接近实时的自服务，无需用户对峰值负载进行工程构造。

⑦ 性能受到监控，同时一致性以及松散耦合架构通过 Web Services 作为系统接口被构建起来。

⑧ 因为数据集中化了，故安全性得到了提升，增加了关注安全的资源等，但对特定敏感数据的失控将是持续关注的，且对内核存储的安全性缺少关注。较传统系统而言，安全性的要求更高。部分原因是提供商可以专注于用户所无法提供的资源之安全性解决方案。然而当“数据分布在更广的范围以及更多数量的设备上”时，以及在由“不相关的多个用户使用的多终端系统“时，安全性的复杂性极大地增加了。用户获取安全审计日志变得不太可能了。私有云的发展动力部分是源自客户对设备的掌控及避免丢失安全信息。

⑨ 维护云计算应用是很简单的，因为显而易见，用户无需再在本机上进行安装。一旦改变达到了客户端，它们将更容易支持以及改进。

1.2.2 云计算简史

云计算的起源要先从互联网演进讲起，图 1-2 所示为云计算的演进与由来。云计算从根本上改变了原有的互联网结构，将计算能力从个人终端向服务端靠拢，弱化了端的概念，提高了计算资源的整体利用率。在量化计算资源的基础上，云计算实现了商业模式由设置向服务进化的过程。更令人满意的是，随着全体物联网的发展，云计算被赋予了更为广泛的定义：从连接计算资源到连接所有人和机器设置，计算能力也将进一步智能化。

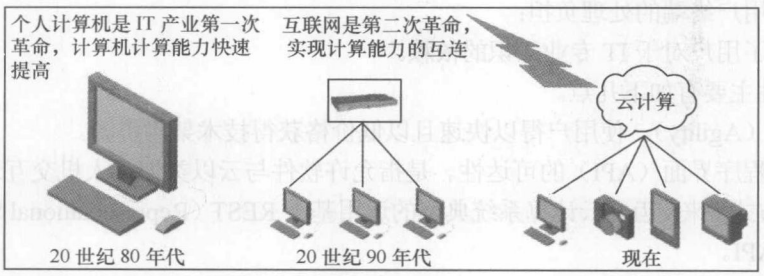


图 1-2 云计算的演进与由来

云计算的发展过程如下。

1983 年，Sun 公司（Sun Microsystems）提出“网络是电脑”（“The Network is the Computer”），2006 年 3 月，亚马逊（Amazon）推出弹性计算云（Elastic Compute Cloud，EC2）服务。

2006 年 8 月 9 日，Google 首席执行官埃里克·施密特（Eric Schmidt）在搜索引擎大会（SES San Jose 2006）首次提出“云计算”（Cloud Computing）的概念。Google“云端计算”源于 Google 工程师克里斯托弗·比希利亚所做的“Google 101”项目。

2007 年 10 月，Google 与 IBM 开始在美国大学校园，包括卡内基美隆大学、麻省理工学院、斯坦福大学、加州大学柏克莱分校及马里兰大学等，推广云计算的计划，这项计划希望能降低分布式计算技术在学术研究方面的成本，并为这些大学提供相关的软硬件设备及技术支持（包括数百台个人电脑及 BladeCenter 与 System x 服务器，这些计算平台将提供 1600 个

处理器，支持包括 Linux、Xen、Hadoop 等开放源代码平台）。

2008 年 1 月 30 日，Google 宣布在中国台湾启动“云计算学术计划”，将与台湾台大、台湾交大等学校合作，将这种先进的大规模、快速计算技术推广到校园。

2008 年 2 月 1 日，IBM 宣布将在中国无锡太湖新城科教产业园为中国的软件公司建立全球第一个云计算中心（Cloud Computing Center）。

2008 年 7 月 29 日，雅虎、惠普和英特尔宣布一项涵盖美国、德国和新加坡的联合研究计划，推出云计算研究测试床，推进云计算。该计划要与合作伙伴创建 6 个数据中心作为研究试验平台，每个数据中心配置 1400 个至 4000 个处理器。这些合作伙伴包括新加坡资讯通信发展管理局、德国卡尔斯鲁厄大学 Steinbuch 计算中心、美国伊利诺伊大学香宾分校、英特尔研究院、惠普实验室和雅虎。

2008 年 8 月 3 日，美国专利商标局网站信息显示，戴尔正在申请“云计算”（Cloud Computing）商标，此举旨在加强这一未来可能重塑的技术。

2010 年 3 月 5 日，Novell 与云安全联盟（CSA）共同宣布一项供应商中立计划，名为“可信任云计算计划（Trusted Cloud Initiative）”。

2010 年 7 月，美国国家航空航天局和包括 Rackspace、AMD、英特尔、戴尔等支持厂商共同宣布 “OpenStack” 开放源代码计划，微软在 2010 年 10 月表示支持 OpenStack 与 Windows Server 2008 R2 的集成；而 Ubuntu 已把 OpenStack 加至 11.04 版本中。

2011 年 2 月，思科系统正式加入 OpenStack，重点研制 OpenStack 的网络服务。

2011 年 10 月 20 日，“盛大云”宣布旗下产品 MongoIC 正式对外开放，这是中国第一家专业的 MongoDB 云服务,也是全球第一家支持数据库恢复的 MongoDB 云服务。

1.2.3　云计算演化

云计算主要经历了四个阶段才发展到现在这样比较成熟的水平，这四个阶段依次是电厂模式、效用计算、网云计算的演进格计算和云计算，演化过程如图 1-3 所示。

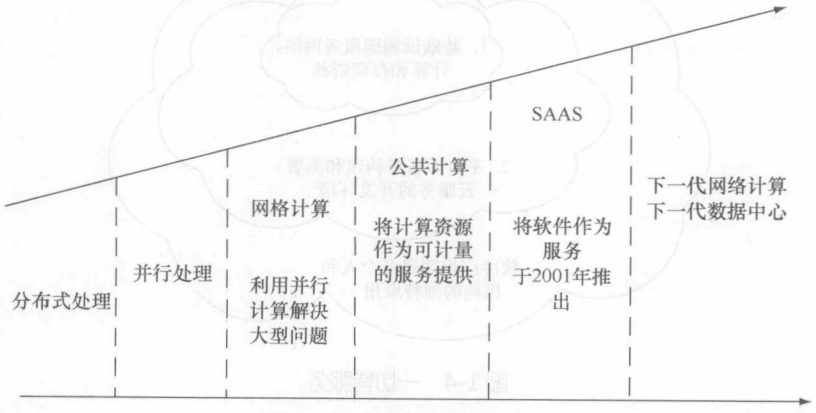


图 1-3　云计算演化过程

（1）电厂模式阶段

电厂模式就好比是利用电厂的规模效应，来降低电力的价格，并让用户使用起来更方便，且无需维护和购买任何发电设备。

（2）效用计算阶段

在 1960 年左右，当时计算设备的价格是非常高昂的，远非普通企业、学校和机构所能承受，所以很多人产生了共享计算资源的想法。1961 年，人工智能之父麦肯锡在一次会议上提出了“效用计算”这个概念，其核心借鉴了电厂模式，具体目标是整合分散在各地的服务器、存储系统及应用程序来共享给多个用户，让用户能够像把灯泡插入灯座一样来使用计算机资源，并且根据其所使用的量来付费。但由于当时整个 IT 产业还处于发展初期，很多强大的技术还未诞生，如互联网等，所以虽然这个想法一直为人称道，但是总体而言“叫好不叫座”。

（3）网格计算阶段

网格计算研究如何把一个需要非常巨大的计算能力才能解决的问题分成许多小的部分，然后把这些部分分配给许多低性能的计算机来处理，最后把这些计算结果综合起来攻克大问题。可惜的是，由于网格计算在商业模式、技术和安全性方面的不足，使得其并没有在工程界和商业界取得预期的成功。

（4）云计算阶段

云计算的核心与效用计算和网格计算非常类似，也是希望 IT 技术能像使用电力那样方便，并且成本低廉。但与效用计算和网格计算不同的是，在需求方面已经有了一定的规模，同时在技术方面也已经基本成熟了。

1.2.4 云服务形式

云的服务形式主要有：交付模式及部署形式。

1. 交付模式

云计算的主要概念是针对于“一切皆服务”术语使用，如图 1-4 所示。

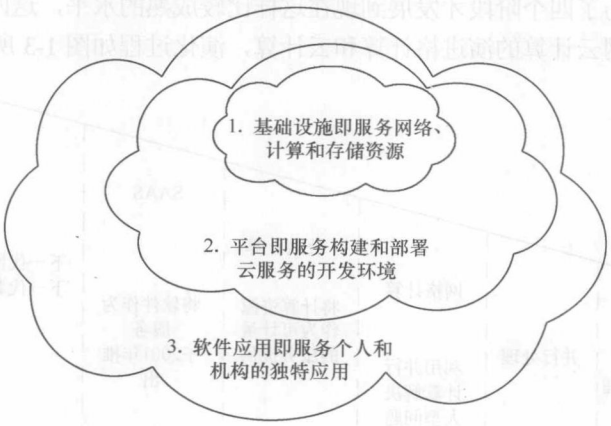


图 1-4 一切皆服务

云计算可描述在从硬件到应用程序的任何传统层级提供的服务如图 1-5 所示。实际上，云服务提供商倾向于提供可分为如下三个类别的服务：把软件当作服务（Software as a Service，SaaS）、把平台当作服务（Platform as a Service，PaaS），以及把基础设施当作服务（Infrastructure as a Service，IaaS）。