



普通高等教育“十三五”规划教材

Hadoop 云技术 从入门到精通

主 编 沈纲祥

副主编 陆冬磊 李泳成 李龙飞



科学出版社

张外信

普通高等教育“十三五”规划教材

Hadoop 云技术从入门到精通

主 编 沈纲祥

副主编 陆冬磊 李泳成 李龙飞

科学出版社

内 容 简 介

本书分 10 章, 全面介绍了 Hadoop 云技术及其在光网络研究方面的真实案例。读者通过本书的学习, 可以掌握搭建及维护 Hadoop 平台的相关知识和技术。

本书可作为高等学校计算机及其相关专业的本科生、研究生大数据课程教材, 也可作为 Hadoop 爱好者以及 Hadoop 企业运维人员的学习参考书。

图书在版编目 (CIP) 数据

Hadoop 云技术从入门到精通 / 沈纲祥主编. —北京: 科学出版社, 2018.3

普通高等教育“十三五”规划教材

ISBN 978-7-03-056618-8

I. ①H… II. ①沈… III. ①数据处理软件-高等学校-教材 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2018) 第 038073 号

责任编辑: 张 帆 / 责任校对: 郭瑞芝
责任印制: 吴兆东 / 封面设计: 迷底书装

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码: 100717

<http://www.sciencep.com>

北京建宏印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 3 月第 一 版 开本: 787×1092 1/16

2018 年 3 月第一次印刷 印张: 10

字数: 250 000

定价: 58.00 元

(如有印装质量问题, 我社负责调换)

前 言

近年来，互联网数据呈现爆炸式增长，数据类型越来越丰富，传统的数据分析技术已不能满足人们对复杂数据信息挖掘的需求。大数据及云计算技术的出现，为人们挖掘数据价值提供了新的方法，已成为当前最热门的技术之一。

本书主要介绍基于 Hadoop 的云计算技术以及相关的一些应用，同时介绍一款小型化的云计算平台 QY1000。全书共 10 章，各章内容安排如下。

第 1 章云计算概述，主要介绍云计算的定义、分类、发展历程和重要技术，为读者提供云计算的一些基础知识。

第 2 章 Hadoop 简介，主要介绍 Hadoop 的起源、发展历程、项目结构、项目介绍及 Hadoop 衍化版本。

第 3 章 MapReduce，主要介绍 MapReduce 架构和工作机制，同时介绍了新一代 MapReduce 框架 YARN。

第 4 章 Hadoop 分布式文件系统，主要介绍 HDFS 的架构、容错机制、备份规则以及权限管理。

第 5 章 Hadoop 案例分析，主要介绍多个具有代表性的利用大数据平台优化公司业务、增强企业竞争力的案例。

第 6 章 Hadoop 集群实验，主要介绍如何利用相关软件搭建大数据实验平台，包括：VMware ESXI 6.5 安装、Ubuntu 16.04 安装、基础运行环境配置、Hadoop 集群部署、HDFS 命令行介绍、Hive 部署、ZooKeeper 和 HBase 部署、Eclipse 开发环境部署等。

第 7 章 Hadoop 系统的优化，主要介绍影响 Hadoop 系统性能的因素、监测工具、如何基于性能基线优化 Hadoop 系统、Map 和 Reduce 的调优以及一些常用的调优手段。

第 8 章 Hadoop 在光网络优化中的应用，主要通过光网络优化设计中的四个问题，验证 Hadoop 云计算平台相对于单机在计算效率上的优势。

第 9 章 Hadoop 的 Java API，主要介绍 Hadoop1.2.1 的 Java API，对几个主要的包、MapReduce 的 API 和 HDFS 的 API 作了详细的说明。

第 10 章 Hadoop 系统微型化和界面可视化，主要介绍自主研发的微型化云计算平台及相应图形化的云平台软件。

本书由苏州大学沈纲祥教授负责总体设计、内容组织和审阅，苏州大学李泳成老

师编写第 1、7、8 章，无锡科技职业学院陆冬磊老师编写第 2、5、6 章，苏州大学李龙飞博士编写第 3、4、9、10 章。衷心感谢苏州大学电子信息学院对本书编写的大力支持。

由于作者水平有限，本书不足之处在所难免，恳请读者提出宝贵意见和建议，以便再版时改进。

编 者

2018 年 1 月

目 录

第 1 章	云计算概述	1
1.1	云计算的定义	1
1.1.1	灵活资源分配	2
1.1.2	按需收费	2
1.1.3	虚拟化	2
1.1.4	网络化	2
1.1.5	高可靠性	3
1.2	云计算的分类	3
1.2.1	按技术路线分类	3
1.2.2	按服务对象分类	4
1.2.3	按资源封装的层次分类	5
1.3	云计算发展简介	6
1.3.1	发展历程	6
1.3.2	国内现状	7
1.3.3	国外现状	7
1.4	云计算技术	8
1.4.1	编程模式	8
1.4.2	虚拟化技术	9
1.4.3	存储技术	9
1.4.4	云计算平台管理技术	10
1.5	小结	10
第 2 章	Hadoop 简介	11
2.1	Hadoop 起源	11
2.2	Hadoop 发展历程	11
2.3	Apache Hadoop 项目及其结构	12
2.4	Hadoop 常用项目介绍	14
2.5	Hadoop 的衍化版本	16
2.5.1	Cloudera	16
2.5.2	Transwarp	17
2.6	小结	17
第 3 章	MapReduce	18

3.1	MapReduce 简介	18
3.2	MapReduce 架构	19
3.3	MapReduce 的工作机制	20
3.3.1	MapReduce 作业运行机制	20
3.3.2	MapReduce 容错机制	21
3.4	MapReduce 编程实例	22
3.5	第二代计算框架 YARN	24
3.5.1	YARN 的设计	24
3.5.2	YARN 的组件	25
3.6	小结	25
第 4 章	Hadoop 分布式文件系统	26
4.1	HDFS 简介	26
4.2	HDFS 架构	26
4.3	HDFS 容错机制	28
4.4	HDFS 备份规则	29
4.4.1	放置策略	29
4.4.2	流水式复制	29
4.5	HDFS 权限管理	29
4.5.1	HDFS 用户身份认证	30
4.5.2	超级用户	30
4.5.3	文件权限管理	30
4.6	小结	31
第 5 章	Hadoop 案例分析	32
5.1	互联网企业——Yahoo、百度、阿里、腾讯	32
5.1.1	Yahoo—Hortonworks	32
5.1.2	百度——百度云	33
5.1.3	阿里巴巴——阿里云	33
5.1.4	腾讯——腾讯云	34
5.2	智慧城市——贵阳	35
5.3	电子商务——京东	36
5.4	国内外旅游业	38
5.4.1	旅游搜索网站 FareCast	38
5.4.2	途牛旅游网	38
5.5	小结	39
第 6 章	Hadoop 集群实验	40
6.1	安装 VMware ESXi 6.5 平台	40
6.2	安装 Ubuntu 16.04 系统	45

6.3 Hadoop 系统运行环境	54
6.3.1 设置主机名	54
6.3.2 设置网络	54
6.3.3 安装 JDK	56
6.3.4 配置系统环境变量	58
6.3.5 配置 SSH 免密登录	59
6.4 Hadoop 集群部署	62
6.4.1 单节点模式 Hadoop 系统安装	63
6.4.2 伪分布模式 Hadoop 系统安装	64
6.4.3 完全分布式 Hadoop 系统安装	65
6.4.4 启动及测试 Hadoop 集群	68
6.5 HDFS 命令行	73
6.6 Hive 部署	74
6.6.1 MySQL 安装	75
6.6.2 Hive 安装	76
6.6.3 启动及测试 Hive	77
6.7 ZooKeeper 和 HBase 部署	77
6.7.1 ZooKeeper 安装	78
6.7.2 启动 ZooKeeper	79
6.7.3 HBase 部署	79
6.7.4 启动 HBase	81
6.8 Eclipse 环境	82
6.8.1 配置工作环境	83
6.8.2 开发初步	89
6.8.3 运行 WordCount 程序	92
6.9 Hadoop 集群运维	94
6.10 小结	94
第 7 章 Hadoop 系统的优化	95
7.1 影响 Hadoop 系统性能的因素	95
7.2 监测工具	96
7.2.1 Chukwa	96
7.2.2 Ganglia 收集器	96
7.2.3 Nagios	96
7.3 基于性能基线的 Hadoop 系统性能优化	97
7.3.1 调优过程	97
7.3.2 甄别资源瓶颈	97
7.4 MapReduce 调优	98

7.4.1	Map 任务调优	98
7.4.2	Reduce 任务调优	99
7.5	常用调优手段	100
7.5.1	面向 MapReduce 应用程序的调优手段	100
7.5.2	面向作业的调优手段	101
7.5.3	面向 Hadoop 系统参数的调优手段	102
7.6	小结	103
第 8 章	Hadoop 在光网络优化中的应用	104
8.1	装箱问题	104
8.2	光网络中的装箱问题	105
8.3	WDM 光网络	106
8.3.1	路由和波长分配	108
8.3.2	绿色光网络的优化设计	110
8.4	弹性光网络	112
8.4.1	基于 SBPP 共享保护策略的路由和频谱分配	113
8.4.2	基于自适应 FEC 分配策略的路由和频谱分配	115
8.5	Hadoop 平台执行多次业务打乱策略	116
8.6	结果分析	117
8.6.1	路由和波长分配	117
8.6.2	绿色光网络的优化设计	118
8.6.3	基于 SBPP 共享保护策略的路由和频谱分配	119
8.6.4	基于自适应 FEC 分配策略的路由和频谱分配	119
8.7	小结	120
第 9 章	Hadoop 的 Java API	121
9.1	Hadoop API 中常用包	121
9.1.1	org.apache.Hadoop.conf 包	121
9.1.2	org.apache.Hadoop.fs 包	122
9.2	MapReduce API	123
9.2.1	Mapper 类	124
9.2.2	Reducer 类	125
9.2.3	Driver	126
9.3	HDFS API	126
9.3.1	获取文件系统	127
9.3.2	创建文件或文件目录	127
9.3.3	删除文件或文件目录	128
9.3.4	写入和读取数据	128
9.3.5	上传和下载文件	128

9.3.6 显示 HDFS 中的文件	129
9.3.7 在 Hadoop 的环境之外配置文件系统	129
9.4 小结	129
第 10 章 Hadoop 系统微型化和界面可视化	130
10.1 微型化	130
10.1.1 平台信息	131
10.1.2 性能	131
10.2 可视化	132
10.2.1 HDFS 文件系统树区域	133
10.2.2 命令控制区域	133
10.2.3 命令行	134
10.2.4 任务状态可视化	135
10.2.5 菜单栏	135
10.2.6 云平台软件的使用	136
10.3 小结	144
附录 缩略语英汉对照表	145
参考文献	148

第 1 章 云计算概述

2006 年 8 月，在搜索引擎大会(SES)上，谷歌公司 CEO 埃里克·施密特(Eric Schmidt)首次提出了云计算(Cloud Computing)的概念。云计算历经多年的发展，如今已成为工业界和学术界研究和应用热点。同时，作为直接影响人类生活的热门技术，云计算的社会关注度极高，包括谷歌(Google)、亚马逊(Amazon)、微软(Microsoft)、阿里巴巴(Alibaba)、腾讯、百度等国内外知名互联网公司纷纷提供商用云计算服务。2016 年，中国信息通信研究院在其发布的《云计算白皮书》中预测：2020 年云计算市场规模将达到 1435.3 亿美元，年复合增长率为 22%(图 1.1)。除美国之外，以中国为代表的云计算新兴国家也处于高速发展之中。

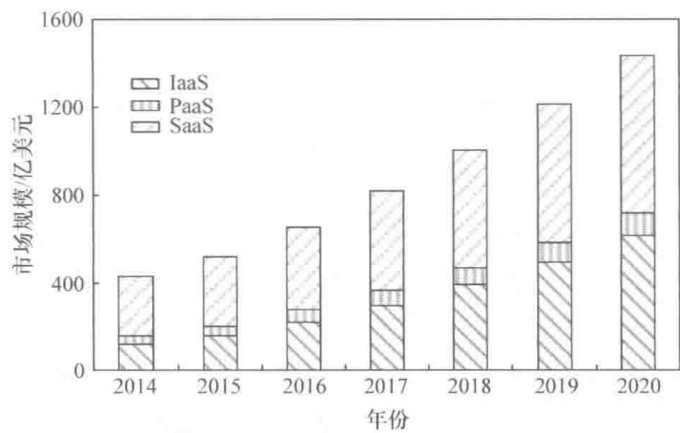


图 1.1 全球云计算市场规模预测

1.1 云计算的定义

研究云计算技术的企业和研究院所众多，它们根据自身研究特点对云计算作了不同的定义。目前云计算的定义仍处于众说纷纭的状态，尚无统一的定义。为人们熟知的云计算定义主要包括以下几种。

- (1) 维基百科：云计算是一种基于互联网的计算方式。在云计算中，共享的软硬件资源和信息可以按需求提供给计算机各种终端和其他设备。
- (2) 美国国家标准与技术研究院：云计算是一种按需付费的模式。在云计算中，可配置的计算资源共享池(资源包括网络、服务器、存储、应用软件服务)可以被便捷地、按需地提供给用户，具有投入管理少、与服务供应商交互少的优势。
- (3) 中国云计算专家咨询委员会：云计算是通过网络提供可伸缩的廉价的分布式计算能力。
- (4) 亚马逊：云计算通过 Internet 云服务平台按需提供计算能力、数据库存储、应用程序和其他 IT 资源，采用按需支付定价模式。

归纳以上比较有代表性的定义后，我们尝试给出自己的见解：云计算是建立于硬件集群之上，具有可动态分配的资源，虚拟化和高可靠性的计算应用服务。

为深入理解云计算，需要分析云计算的一些基本特征。与传统的计算模式相比，云计算具有以下五个主要特性：灵活资源分配、按需收费、虚拟化、网络化、高可靠性。接下来我们详细介绍每种特性。

1.1.1 灵活资源分配

在云计算的定义中存在资源池的概念。资源池将云计算平台中所有的计算和存储资源集中，统一管理，并能灵活地按需分配给用户。资源池是云计算最为重要的组成部分。在云计算中，资源池具备两个重要特性。

(1) 针对不同的用户需求，资源池可以灵活地为用户分配所需的资源。

(2) 针对同一用户需求，资源池也可以动态地增减所提供的资源，以满足用户实时变化的资源需求。

资源池保证了云计算平台的计算和存储资源都是按照每一个用户的实际需求提供，这极大地提高了平台的资源利用率；同时，对于用户来说，他们无须额外购买高性能的硬件设备来满足其变化的资源需求，所以可以极大地节省其成本。

1.1.2 按需收费

资源池能够根据用户需求灵活地分配资源，因此用户只需为其所占用的资源缴纳费用即可。在用户的资源需求实时变化的情况下，云计算平台也只是按照实际使用的资源进行收费，并非按照峰值收费或者按使用次数收费，这样的收费模式称为按需收费模式。在通常情况下，资源池的容量远远大于用户所需的资源量，这就使得云计算应用的开发者可以在理论上使用无限大的资源。这给用户带来了极大的便利，使开发者无须考虑资源的限制，有更大的开发空间。当然，由于按需收费，开发者自身会从减少资源的使用量、节约成本的角度对资源的使用进行优化。

1.1.3 虚拟化

虚拟化是指采用虚拟化技术将一台计算机模拟为多台虚拟计算机。每台虚拟计算机可以采用不同的操作系统，运行不同的应用程序，且互不影响。虚拟化技术的最大优势在于可以显著提高计算机的使用效率。具体到云计算中，虚拟化技术是实现云计算平台硬件资源的虚拟化管理、分配和使用的基础。通过虚拟化技术，云计算平台可以为用户分配独立的计算、存储和数据资源等，用户可以通过网络在本地计算机上透明地使用这些资源。

1.1.4 网络化

云计算平台一般是通过网络提供服务的。基于云计算平台开发的应用，都需将云计算数据中心的计算和存储资源封装，并通过网络为用户提供服务。云计算技术必须支持资源的网络化接入，这样才能有效地向应用开发者和用户提供服务，这种依赖于网络的

特性称为网络化。网络是云计算平台所必须依赖的信息传输媒介，没有网络，云计算平台将无用武之地。因此，网络技术的发展是促进云计算技术发展的重要基础之一。

1.1.5 高可靠性

云计算平台依赖大量的底层硬件设备，然而任何设备都会出现异常。一旦受到不可抗力因素(如火灾、地震、战争等)的影响，底层的硬件设备可能会发生故障，这些都可能造成大量存储在云计算平台上的数据丢失，从而造成极大的影响。为防止以上情况发生，云计算平台通常都提供防灾备份功能，以保证其高可靠性。用户数据及服务被备份到不同的服务器上，一旦某个节点出现故障，云计算平台承载的业务会被自动切换到备份节点上，保证所运行的应用不受影响。为了实现防灾备份，云计算平台的节点在地理上也是分布式的，以免在出现大范围灾难时同时被毁，导致用户数据不可恢复。

1.2 云计算的分类

常见的云计算分类方式有以下三种：①基于技术路线，分为资源整合型和资源切分型；②基于服务对象，分为公有云、私有云和混合云；③基于资源的封装层次，分为基础设施即服务、平台即服务和软件即服务，如表 1.1 所示。本节将详细介绍各种类型的云计算。

表 1.1 云计算分类

分类方式	云计算类型
技术性路线	资源整合型、资源切分型
服务对象	公有云、私有云、混合云
资源的封装层次	基础设施即服务、平台即服务、软件即服务

1.2.1 按技术路线分类

资源整合型云计算通过构建集群架构，整合大量节点的计算资源和存储资源来提供云服务。资源整合型云计算建立了跨节点、弹性化的资源池，其核心技术包括分布式计算和存储技术。常见的资源整合型云计算系统包括 MPI、Hadoop、HPCC 等。

资源切分型云计算通过虚拟化技术对服务器进行弹性切分，从而有效地提高服务器的利用率。资源切分型云计算的关键技术是虚拟化。由于具有很高的资源利用率，这类云计算也是目前应用较为广泛的技术，但是它也存在跨节点资源整合能力低的缺点。两种类型的核心技术及实例如表 1.2 所示

表 1.2 资源整合型和资源切分型云计算

类型	核心技术	实例
资源整合型	分布式计算和存储技术	MPI、Hadoop、HPCC
资源切分型	虚拟化	KVM、VMware

1.2.2 按服务对象分类

按服务对象分类，云计算可以分为公有云、私有云和混合云。顾名思义，公有云是面向所有用户开放的云计算服务。任何个人或企业只需缴纳一定的费用，即可使用云计算服务商提供的计算和存储资源。在日常生活中，比较典型的云盘服务商，如 360 云盘(已关闭)、百度云盘、Dropbox 等，都是公有云的代表，它们提供云存储服务。此外，包括阿里巴巴、谷歌、亚马逊等也纷纷提出了更为全面稳定的公有云服务。这些服务商为用户提供包括应用程序、软件运行环境等各种各样的云计算和存储资源。然而，公有云在安全性上存在不足。一些重要的信息和程序一旦进入公有云，所有者将失去完全控制权，无法保证不被他人窃取。因此，具有重要保密级别的信息不应上传到公有云上。

由于公有云存在安全性的缺陷，部分企业自发地构建了只为本公司服务的云计算平台，此类主要服务于某个组织内部的云计算平台称为私有云。私有云具备公有云的功能和特性。出于安全性的考虑，私有云并不对公众开放。相比于公有云的信息不可控性，私有云的高信息可控性使其非常适合于具有重要机密的大型企业和政府部门，以及保密级别极高的军事单位。

虽然私有云能够确保机密信息不被外泄，但很多企业依然需要大量的数据存储和计算能力来处理其非关键信息。企业如果自己建立一个非常强大的数据中心，需要花费大量的人力和财力，即使对于一个大型企业来说，这也是非常不明智的。因此，结合公有云和私有云的特点，人们提出了混合云。混合云有效地解决了公有云的安全性问题和私有云的成本问题，兼有公有云和私有云的优点。基于混合云，用户可以租用公有云服务，利用公有云强大的计算能力承载公司的部分业务，并将非关键信息存储到公有云上。同时，企业付出比较低的代价建立自己的私有云，将关键信息和服务放到私有云上。当然，如何在公有云和私有云之间实现互连互通是混合云需要解决的难点。

目前，公有云占据了最大的市场份额，这主要受益于其“按需付费”的商业模式，它允许所有不同类型的用户使用云服务，从而保证了其市场规模。其次，由于私有云具有更高的安全性，私有云市场也不容忽视，仅次于公有云。相比于公有云的成熟和私有云的安全，混合云继承了两者的优势，并对彼此的弱势进行了互补。虽然混合云因为技术、维护等原因目前还不是市场主流，但其未来的发展空间是巨大的。公有云、私有云和混合云优缺点对比如表 1.3 所示。

表 1.3 按服务对象分类

类型	优点	缺点
公有云	用户成本低	安全性差
私有云	安全性高	用户成本高
混合云	兼顾成本和安全性	技术难度较高

1.2.3 按资源封装的层次分类

1. 基础设施即服务(IaaS)

IaaS 指的是云计算平台把计算和存储资源未经封装，直接通过网络，以服务的形式提供给用户使用。在云计算平台中，用户无法掌控云计算平台的架构，但可以使用基础计算资源和存储资源。类似于大型基础设施的供应商(如电厂、电信运营商)，云计算服务商只需要按照用户的需求提供相应的资源即可。在此模式下，用户无须购买硬件设备和聘请大量的运维人员，从而可以节省大量的资金，有利于企业集中精力发展其主业。此类云服务的对象一般为具有较强专业知识和能力的资源使用者。

2. 平台即服务(PaaS)

PaaS 指的是云计算平台把计算和存储资源封装后，以某种接口和协议的形式提供给用户调用，资源的使用者不再面向底层要资源，即资源的使用者不需要管理或者控制底层的云基础设施，包括网络、服务器、操作系统、存储等，但客户能控制部署的应用程序，也可以配置运行应用程序的托管环境。平台即服务位于云计算的中间层，主要面向软件开发者或者软件开发商。这样的云平台往往会向开发者提供大量的 API 或者插件，软件开发人员可以直接在云端完成程序的编写、编译和运行，并将自己的程序在云平台上托管。

3. 软件即服务(SaaS)

SaaS 指的是云计算平台将计算和存储资源封装成用户可以直接使用的软件，并通过互联网提供给用户。SaaS 面向的服务对象为一般终端用户，用户只使用软件，无须了解任何云计算系统的内部结构，也不需要具有专业的软件开发技能。软件即服务是一种以软件服务为理念，为用户提供商用软件版权的成熟商业模式。

以上三种服务对应的服务体系结构如图 1.2 所示。

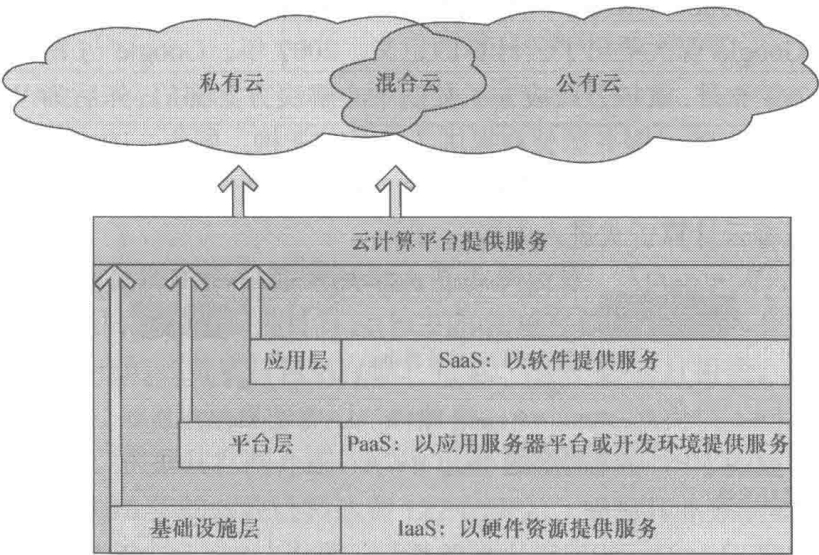


图 1.2 云计算服务体系结构

表 1.4 给出了三种不同服务类型的云服务产品。

表 1.4 不同服务类型的云服务产品

按资源封装层次分类	云计算产品
IaaS	Amazon EC2、IBM BlueCloud、Rackspace
PaaS	IBM IT Factory、Google APP Engine、Force.com
SaaS	Google APPS、SoftWare+Service、Microsoft CRM

1.3 云计算发展简介

云计算被认为是科技界的又一次革命性的发明，它将改变人们的工作和生活方式。云计算的出现与并行计算、分布式计算、网格计算等有着千丝万缕的关系。

1.3.1 发展历程

图 1.3 给出了云计算技术发展的主要历程。1959 年，虚拟化技术的提出为云计算奠定了基础。1984 年，Sun 公司联合创始人提出“网络即计算机”，被认为是对云计算的早期构想，具有重要的历史意义。1994 年和 1995 年，网格计算和并行计算相继被提出，并建立了实验平台。1998 年，著名的 VMware 公司成立并首次引入了基于 X86 的虚拟化技术。1999 年，第一个商业化的 IaaS 平台建成。2000 年，SaaS 兴起。

2004 年，Google 首次发布采用 MapReduce 模型进行分布式编程的论文。之后，基于 MapReduce 模型，Hadoop 开源项目启动，其主要由 HDFS、MapReduce 和 HBase 组成。其中，分布式文件系统 HDFS 由 Google File System(GFS)演进而来；MapReduce 由 Google 的 MapReduce 开源实现；HBase 由 Google 的 BigTable 开源实现。由于 HDFS 和 MapReduce 的强大功能，Hadoop 日后成为最具代表性的分布式系统基础架构。

2006 年，Google 首次给出了云计算的定义。2007 年，Google 与 IBM 在大学开设云计算课程。2007 年 3 月，戴尔公司成立了数据中心解决方案部门，先后为 Windows Azure、Facebook 和 Ask.com 等云服务运营商提供了云基础架构。同年，IBM 也首次发布了云计算商业解决方案，推出了 BlueCloud 计划。2008 年，IBM 在中国无锡建立了国内首个云计算中心，标志着云计算正式进入中国。

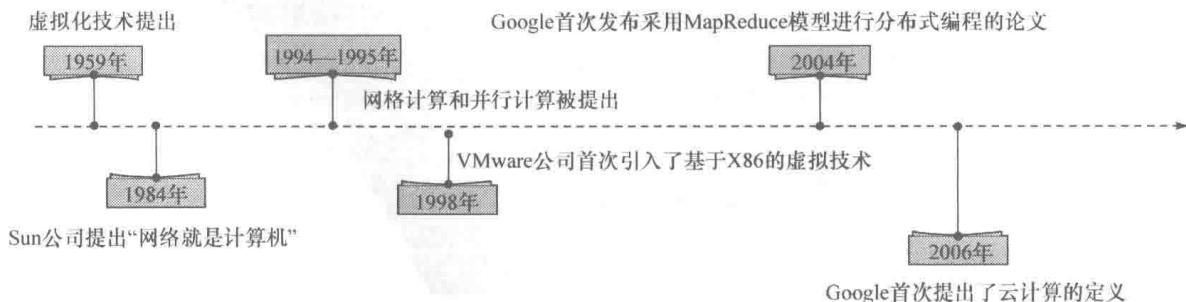


图 1.3 云计算技术发展时间轴

1.3.2 国内现状

国内较早提出私有云解决方案的公司是华胜天成，其“揽胜行动”包括 IaaS 管理平台、华胜天成蓝云管理平台、绿色机房、虚拟桌面等基于云计算的服务平台。华为公司也提供了云计算服务，主要是针对企业的通用云计算解决方案，包括通信、会议、医疗、交通、金融等领域。如今，华为的云计算业务已遍布全球，包括为海外运营商沃达丰开发了云运营平台、为巴西建立了云呼叫中心平台等。除企业之外，我国政府对云计算也十分重视。在“十二五”规划中，明确将云计算技术作为信息化产业发展的重点领域之一，这促使了地方政府积极发展云计算市场。在 2008 年，无锡市政府与 IBM 合作，建立了私有云“云谷”，并帮助包括七匹狼等公司开展电子商务。2011 年 4 月，重庆市建立了国内最大的云计算试验区。另外，上海市政府的部分单位分别与东软集团、浪潮软件、万达信息在智能交通、电子政务、卫生健康等领域达成了云计算应用合作协议。

中国信息通信研究院公布的数据显示，2015 年中国公共云服务市场整体规模达到 102.4 亿元，比 2014 年增长 45.8%。根据 IDC 发布的 2015 年中国公有云计算报告，阿里云的营收达 2.59 亿美元，占 31% 的市场份额，排名第一，其次分别为中国电信、中国联通、世纪互联(微软云授权实体)等。阿里云的业务体量很快超过 Google 云计算的全球体量，进入全球前三，形成“3A”(亚马逊 AWS、微软 Azure、Alibaba Cloud)竞争格局。在国内，除了阿里巴巴之外，还有包括腾讯、华为、百度等服务提供商。这些企业将在公有云、私有云和混合云等所有的云计算细分市场里和阿里巴巴竞争。表 1.5 列出了我国云计算产业各个领域的主要服务商。

表 1.5 国内主要云计算服务商

领域	主要服务商
IT 基础设施与系统集成服务	浪潮信息、华胜天成、浙大网新等
IaaS	中国电信、中国联通、中国移动等
PaaS	阿里巴巴、华为等
SaaS	阿里巴巴、三五互联、焦点科技等

1.3.3 国外现状

国外主要的云计算服务商包括亚马逊、谷歌和微软。Amazon 的云计算业务称为 AWS，于 2006 年推出，采用弹性计算云(EC2)和简单存储服务(S3)为企业提供计算与存储服务。Amazon 现有的收费项目包括存储、带宽、CPU 以及月租费。其中，存储服务器和带宽按容量收费，CPU 则根据使用时长收费。Amazon 也提供不同的套餐供用户选择，价格也不尽相同。根据其官网介绍，AWS 已经为全球 190 个国家/地区成百上千家企业提供服务。AWS 2016 年的营收已经达到 122 亿美元，成为其业务中最重要的一部分。

相比于 Amazon，Google 被认为是云计算的最大使用者。Google 搜索引擎本身就建立在分布于 200 多个不同地点、超过 100 万台服务器的集群之上，这些设备的数量正在