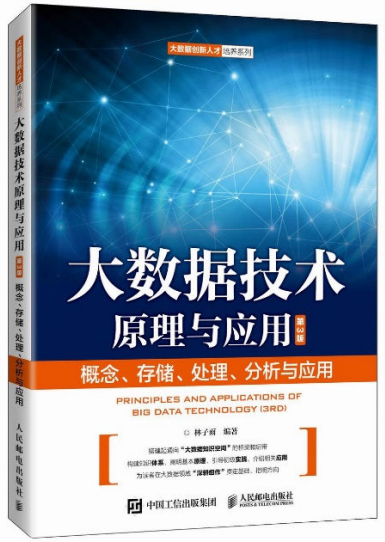
大数据技术原理与应用

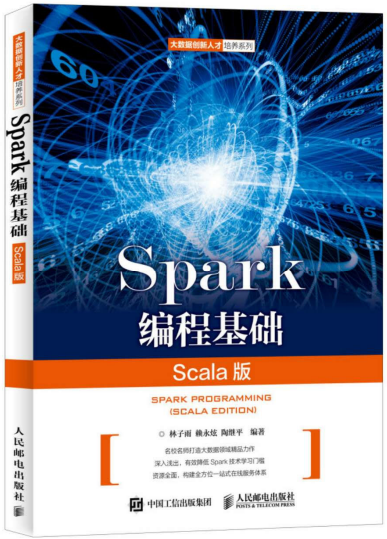
1. 大数据概述

陈建文
电子信息与通信学院
chenjw@hust.edu.cn

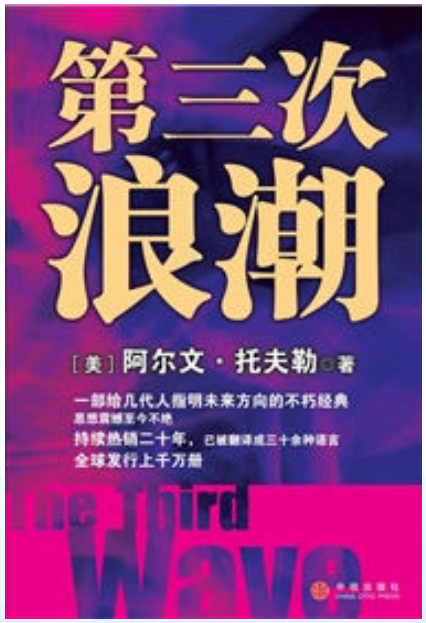
1. 大数据概述

1.1 大数据时代
1.2 大数据概念
1.3 大数据的影响
1.4 大数据的应用
1.5 大数据关键技术
1.6 大数据计算模式
1.7 大数据产业
1.8 大数据与云计算物联网关系



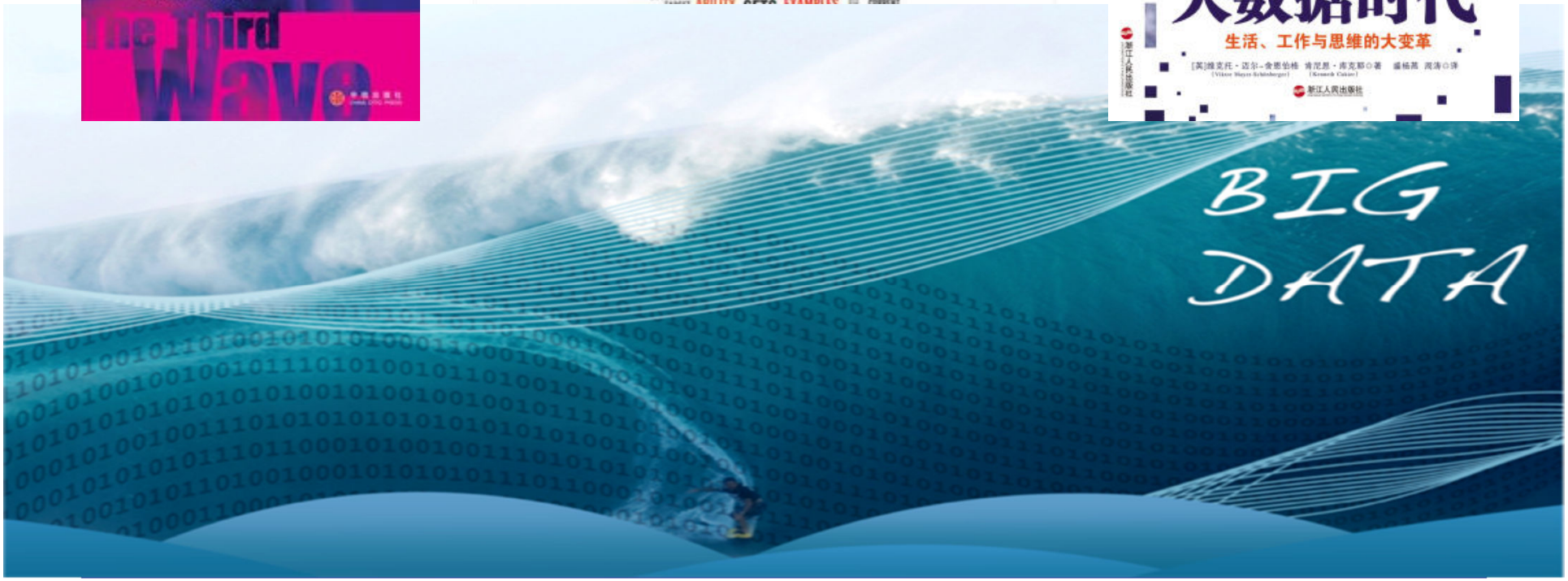


1.1 大数据时代


第三次浪潮
[美] 阿尔文·托夫勒 著
一部给几代人指明未来方向的不朽经典
持续热销二十年，已被翻译成三十余种语言
全球发行上千万册
The Third Wave


BIG
DATA
ANALYTICS
SOFTWARE
SIZE
NEEDED
PETABYTES
INTERNET
TECHNOLOGIES
MANAGEMENT
SOCIAL
LARGE
EVERY
STORAGE
RESEARCH
DATABASES
COMPLEX
INFORMATION
NETWORKS
INCLUDE
SYSTEMS
TIME
SHARED
DISK
HUNDREDS
SEARCH
RECORDS


大数据时代
BIG
DATA
A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK
大数据时代
生活、工作与思维的大变革
浙江人民出版社


BIG
DATA

1.1.1 第三次信息化浪潮

表1-1 三次信息化浪潮

信息化浪潮	发生时间	标志	解决问题	代表企业
第一次浪潮	1980年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010年前后	物联网、云计算和大数据	信息爆炸	将涌现出一批新的市场标杆企业

根据IBM前首席执行官郭士纳的观点，IT领域每隔十五年就会迎来一次重大变革！！！

1.1.2 信息科技为大数据时代提供技术支撑

1. 存储设备容量不断增加

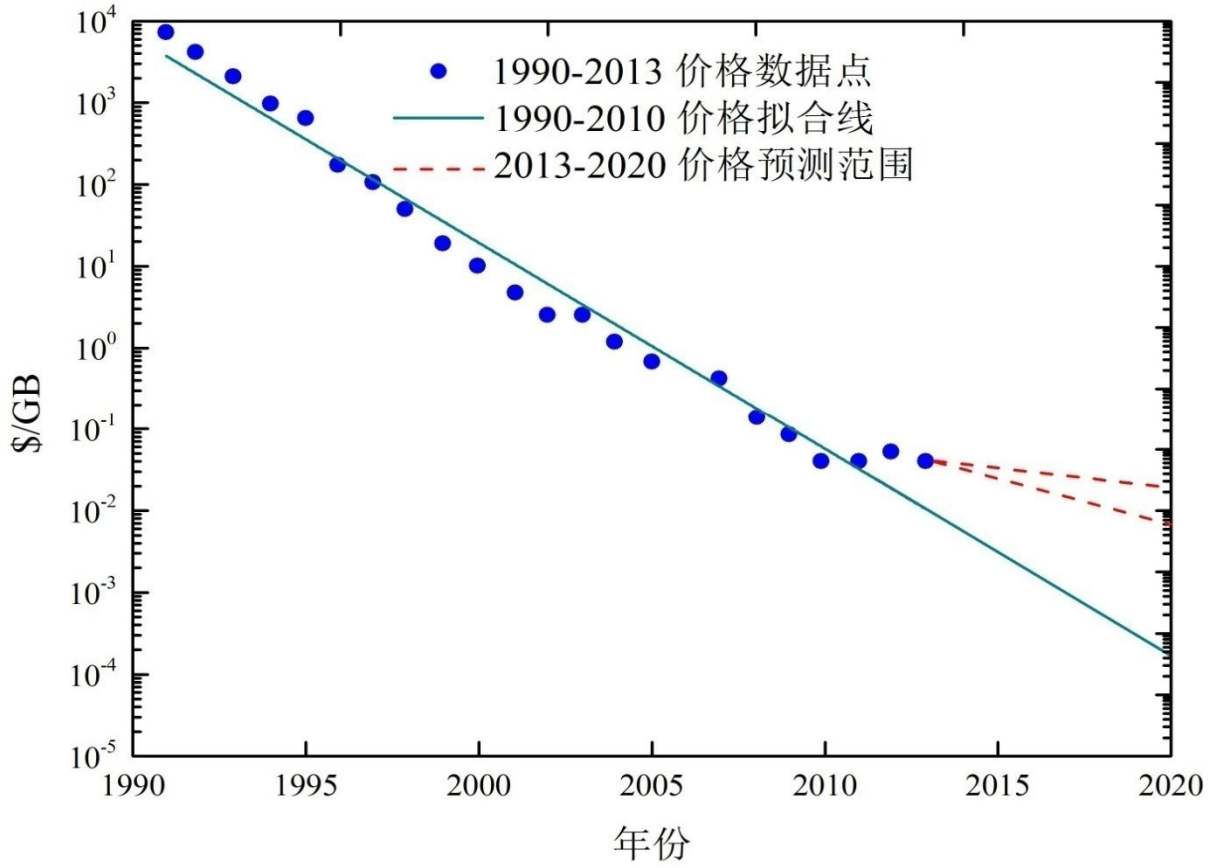


图1-1 存储价格随时间变化情况

2. CPU处理能力大幅提升

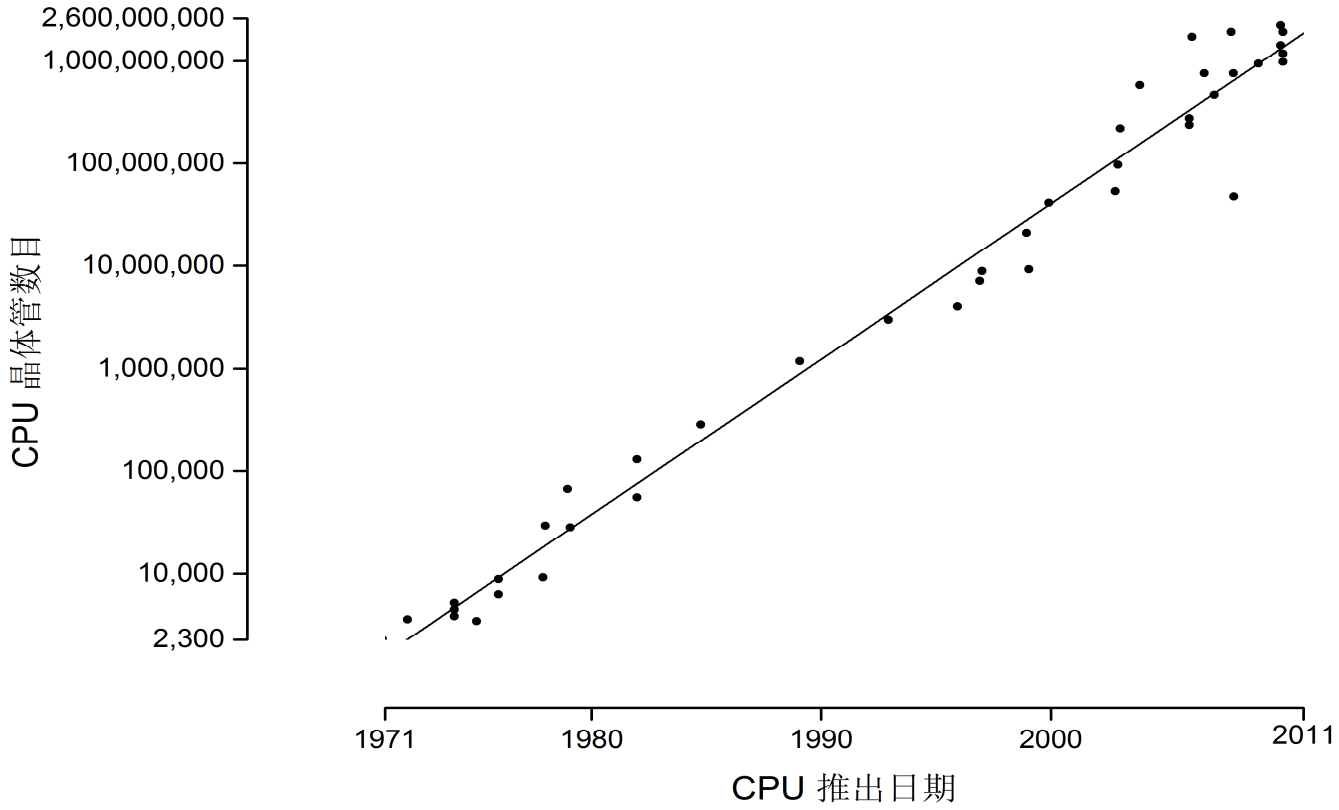


图1-3 CPU晶体管数目随时间变化情况

3. 网络带宽不断增加

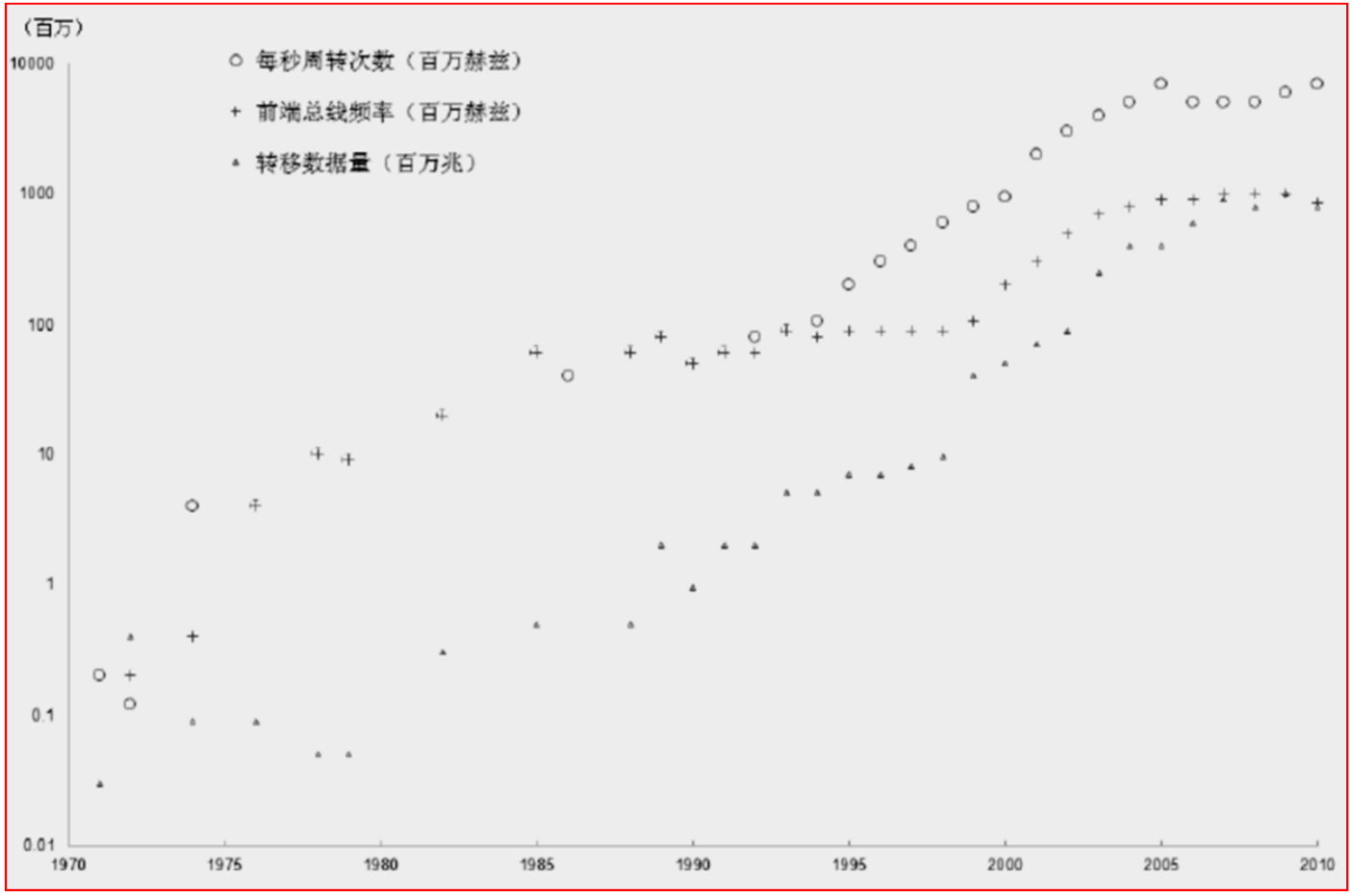


图1-4 网络带宽随时间变化情况

在信息化基础设施方面，据工业和信息化部官网消息，截至2019年12月底，我国互联网宽带接入端口数量达9.16亿个，其中，光纤接入端口占互联网接入端口的比重达91.3%；光缆线路总长度已达4750万公里，相当于在京沪高铁线上往返1.8万余次。同时，近五年来固定宽带和移动宽带资费平均下降90%，速率提升6倍。目前，我国已基本实现“城市光纤到楼入户，农村宽带进乡入村”。

据中国信息通信研究院（简称中国信通院）数据，截至2020年2月底，全国建设开通5G基站达16.4万个，5G网络建设基础不断夯实。2020年中国将建设60万～80万个5G基站。

1.1.3 数据产生的变革促成大数据时代的来临

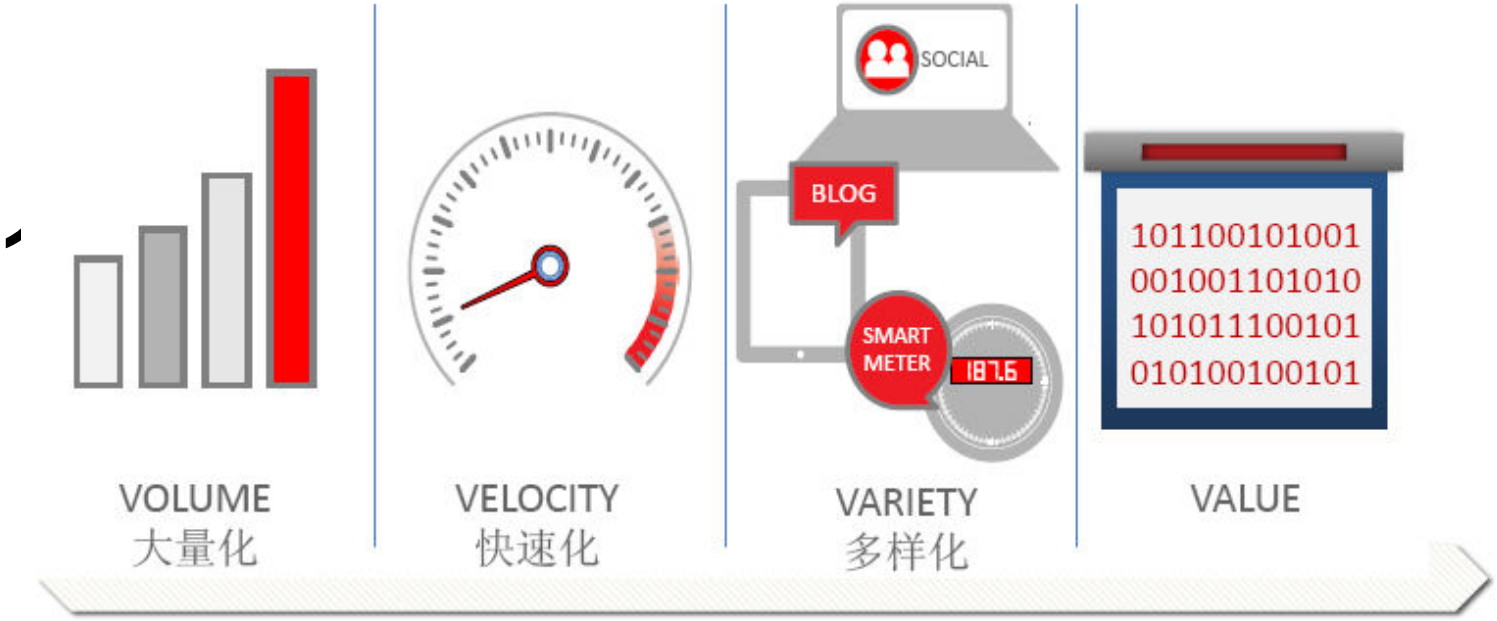
图1-5 数据产生方式的变革

1.1.4 大数据的发展历程

表1-2 大数据发展的三个阶段

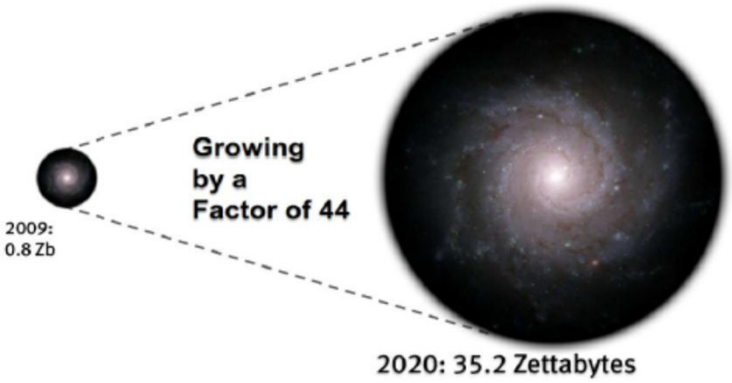
阶段	时间	内容
第一阶段：萌芽期	上世纪90年代至本世纪初	随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等。
第二阶段：成熟期	本世纪前十年	Web2.0应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心技术，谷歌的GFS和MapReduce等大数据技术受到追捧，Hadoop平台开始大行其道。
第三阶段：大规模应用期	2010年以后	大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅提高。

1.2 大数据概念



大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

1.2.1 大数据量



TERABYTE	10 的 12 次方	一块 1TB 硬盘		200,000 照片或 mp3 歌曲
PETABYTE	10 的 15 次方	两个数据中心机柜		16 个 Blackblaze pod 存储单元
EXABYTE	10 的 18 次方	2,000 个机柜		占据一个街区的 4 层数据中心
ZETTABYTE	10 的 21 次方	1000 个数据中心		纽约曼哈顿的 1/5 区域
YOTTABYTE	10 的 24 次方	一百万个数据中心		特拉华州和罗德岛州

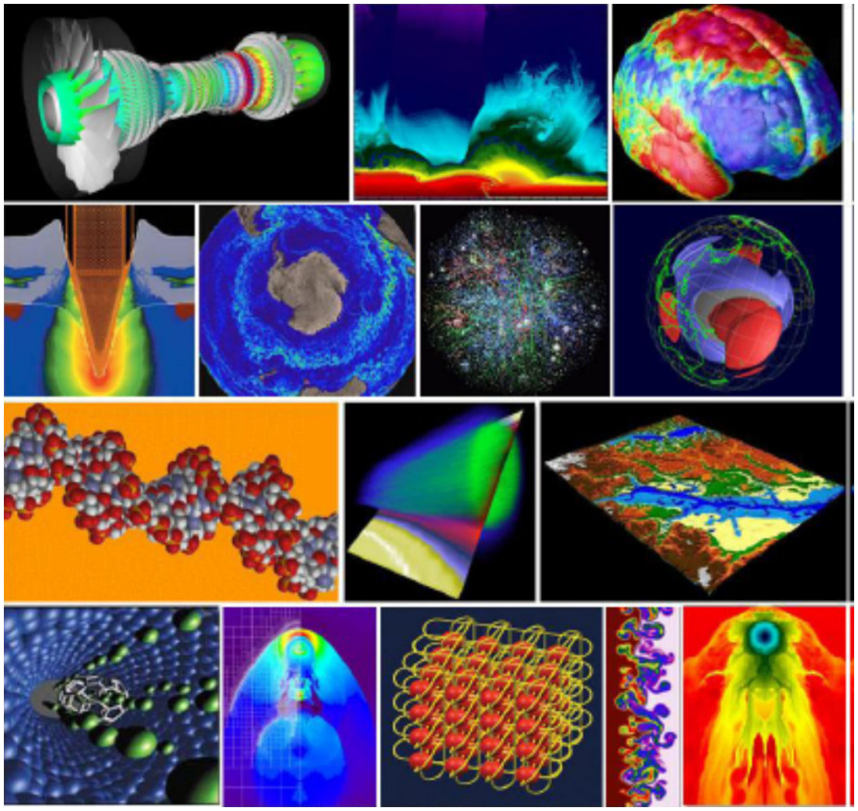
■ 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）；

■ 人类在最近两年产生的数据量相当于之前产生的全部数据量；

■ 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍。

1.2.2 数据类型多

- 大数据是由结构化和非结构化数据组成的
  - 10%的结构化数据，存储在数据库中
  - 90%的非结构化数据，它们与人类信息密切相关



- 科学研究
  - –基因组
  - –LHC 加速器
  - –地球与空间探测
- 企业应用
  - –Email、文档、文件
  - –应用日志
  - –交易记录
- Web 1.0数据
  - –文本
  - –图像
  - –视频
- Web 2.0数据
  - –查询日志/点击流
  - –Twitter/ Blog / SNS
  - –Wiki

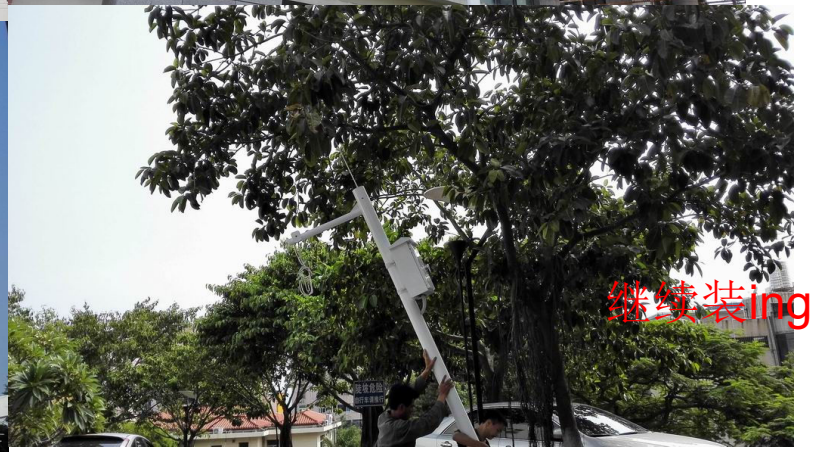
1.2.3 处理速度快

- 从数据的生成到消耗，时间窗口非常小，可用于生成决策的时间非常少；
- 1秒定律：这一点也是和传统的数据挖掘技术有着本质的不同。

1.2.4 价值密度低

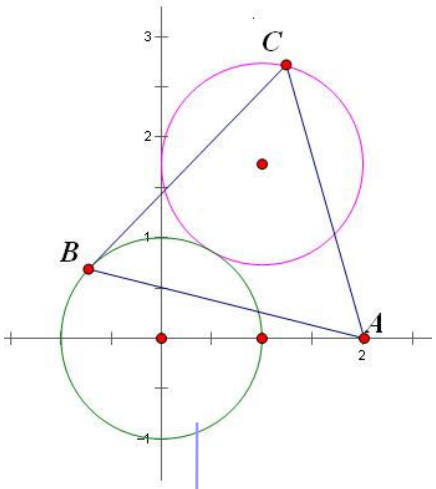
价值密度低，商业价值高。

以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的商业价值。



1.3 大数据的影响

图灵奖获得者、著名数据库专家Jim Gray 博士观察并总结人类自古以来，在科学研究上，先后历经了实验、理论、计算和数据四种范式。



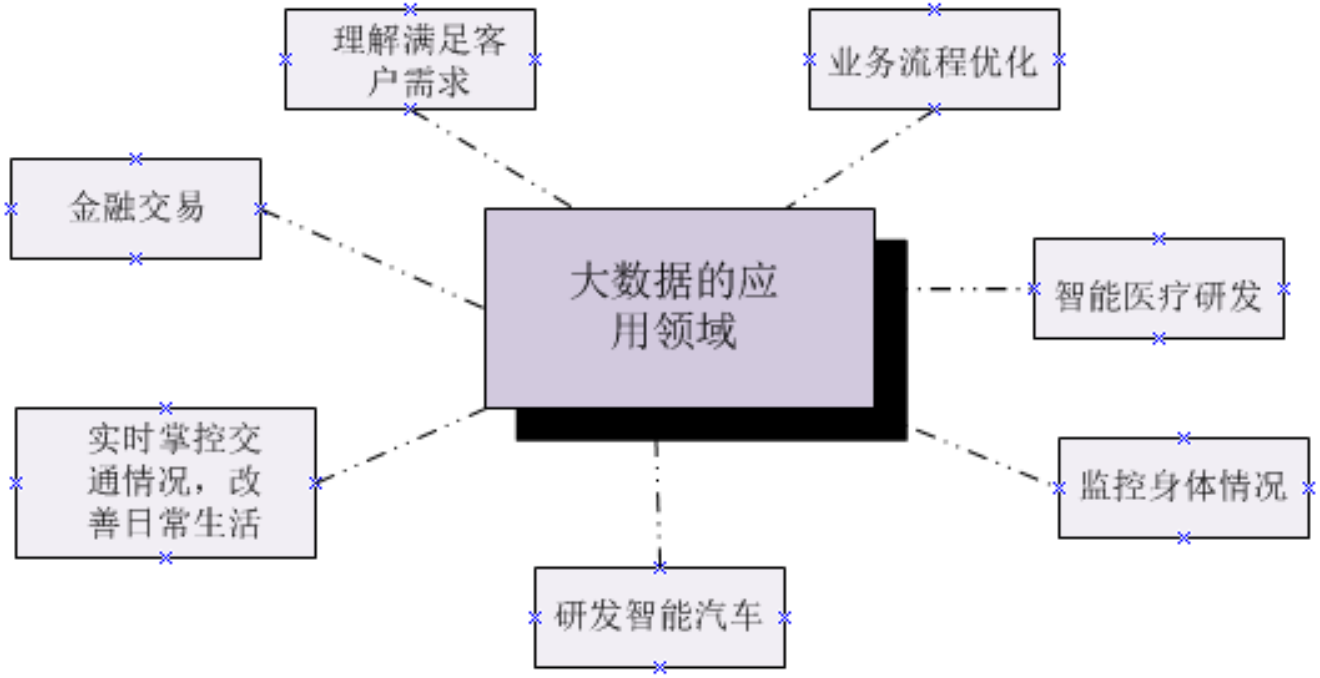
实验　　理论　　计算　　数据

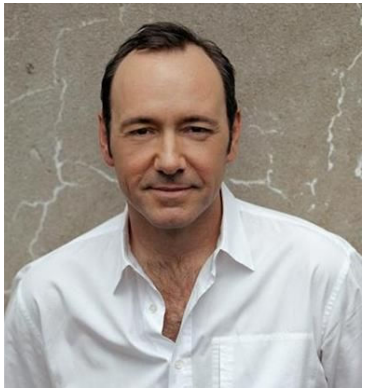
- 在思维方式方面，大数据完全颠覆了传统的思维方式：
  - 全样而非抽样
  - 效率而非精确
  - 相关而非因果

- 在社会发展方面，大数据决策逐渐成为一种新的决策方式，大数据应用有力促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现；
- 在就业市场方面，大数据的兴起使得数据科学家成为热门职业；
- 在人才培养方面，大数据的兴起，将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制；

1.4 大数据的应用

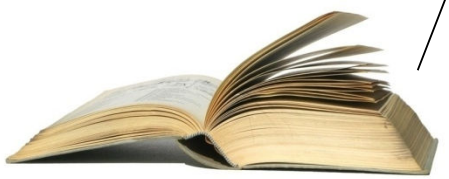
- 大数据无处不在，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业都已经融入了大数据的印迹。



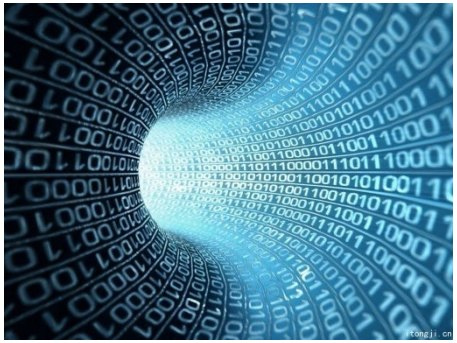


Kevin Spacey

David Fincher



英国同名小说《纸牌屋》

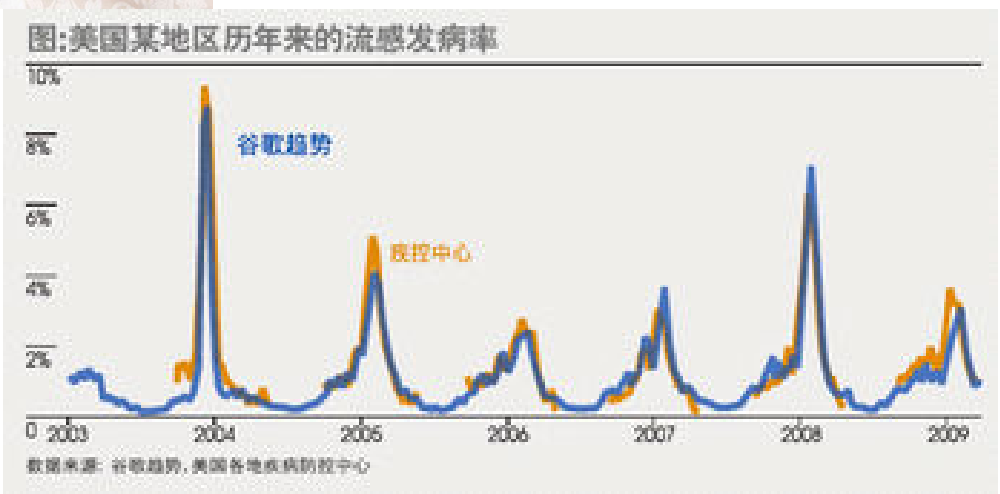


大数据分析

风靡全球的美剧《纸牌屋》

从谷歌流感趋势看大数据的应用价值

“谷歌流感趋势”，通过跟踪搜索词相关数据来判断全美地区的流感情况

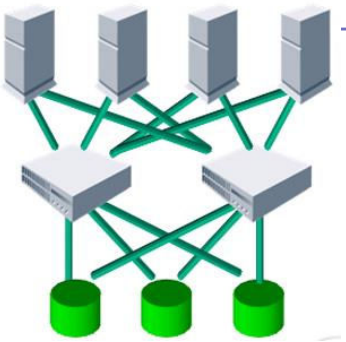


1.5 大数据关键技术

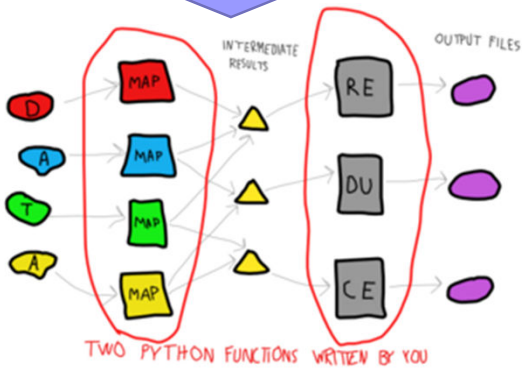
表1-3 大数据技术的不同层面及其功能

技术层面	功能
数据采集	利用ETL工具将分布的、异构数据源中的数据如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础；或者也可以把实时采集的数据作为流计算系统的输入，进行实时处理分析。
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理。
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析；对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据。
数据隐私和安全	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全。

两大核心技术



分布式存储



分布式处理

GFS\HDFS
BigTable\HBase
NoSQL（键值、列族、图形、文档数据库）
NewSQL（如：SQL Azure）

MapReduce

1.6 大数据计算模式

表1-4 大数据计算模式及其代表产品

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala等

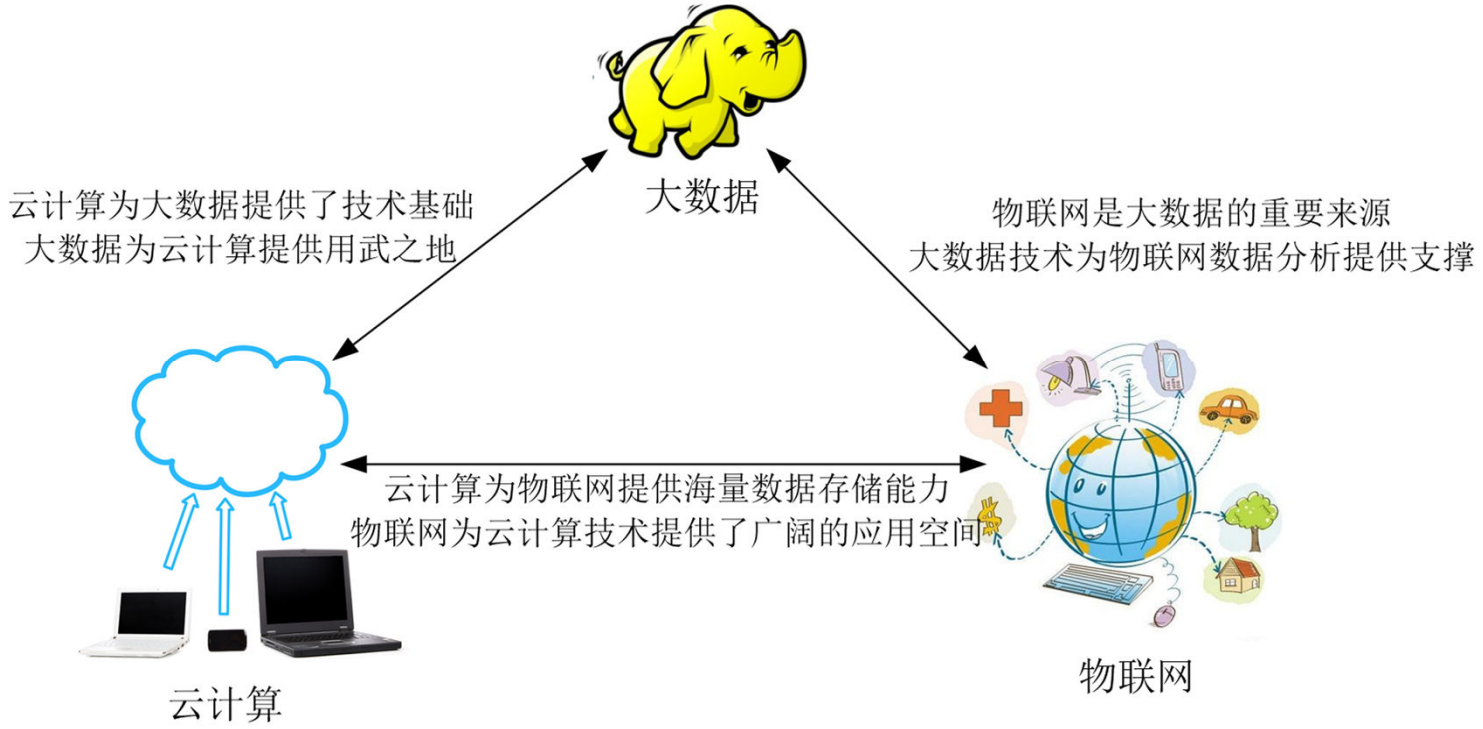
1.7 大数据产业

- 大数据产业是指一切与支撑大数据组织管理和价值发现相关的企业经济活动的集合。

产业链环节	包含内容
IT基础设施层	包括提供硬件、软件、网络等基础设施以及提供咨询、规划和系统集成服务的企业，比如，提供数据中心解决方案的IBM、惠普和戴尔等，提供存储解决方案的EMC，提供虚拟化管理软件的微软、思杰、SUN、Redhat等
数据源层	大数据生态圈里的数据提供者，是生物大数据（生物信息学领域的各类研究机构）、交通大数据（交通主管部门）、医疗大数据（各大医院、体检机构）、政务大数据（政府部门）、电商大数据（淘宝、天猫、苏宁云商、京东等电商）、社交网络大数据（微博、微信、人人网等）、搜索引擎大数据（百度、谷歌等）等各种数据的来源
数据管理层	包括数据抽取、转换、存储和管理等服务的各类企业或产品，比如分布式文件系统（如Hadoop的HDFS和谷歌的GFS）、ETL工具（Informatica、Datastage、Kettle等）、数据库和数据仓库（Oracle、MySQL、SQL Server、HBase、GreenPlum等）
数据分析层	包括提供分布式计算、数据挖掘、统计分析等服务的各类企业或产品，比如，分布式计算框架MapReduce、统计分析软件SPSS和SAS、数据挖掘工具Weka、数据可视化工具Tableau、BI工具（MicroStrategy、Cognos、BO）等等
数据平台层	包括提供数据分享平台、数据分析平台、数据租售平台等服务的企业或产品，比如阿里巴巴、谷歌、中国电信、百度等
数据应用层	提供智能交通、智慧医疗、智能物流、智能电网等行业应用的企业、机构或政府部门，比如交通主管部门、各大医疗机构、菜鸟网络、国家电网等

1.8 大数据与云计算物联网关系

云计算、大数据和物联网代表了IT领域最新的技术发展趋势，三者既有区别又有联系。



1.8.1 云计算

1. 云计算概念

- 云计算实现了通过网络提供可伸缩的、廉价的分布式计算能力，用户只需要在具备网络接入条件的地方，就可以随时随地获得所需的各种IT资源.

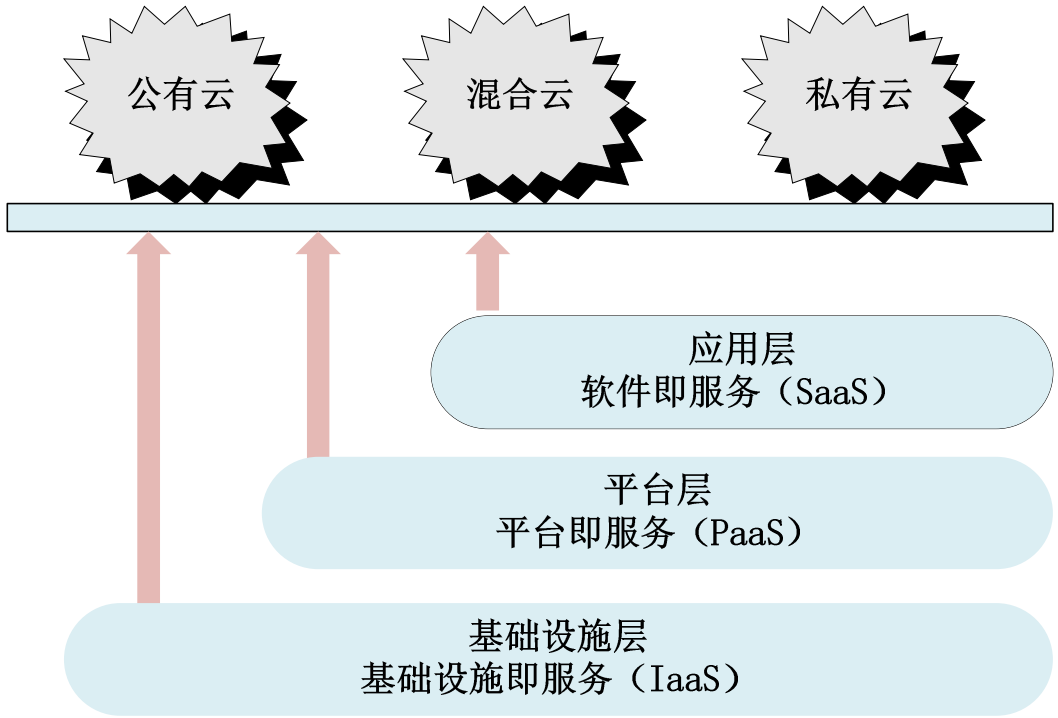
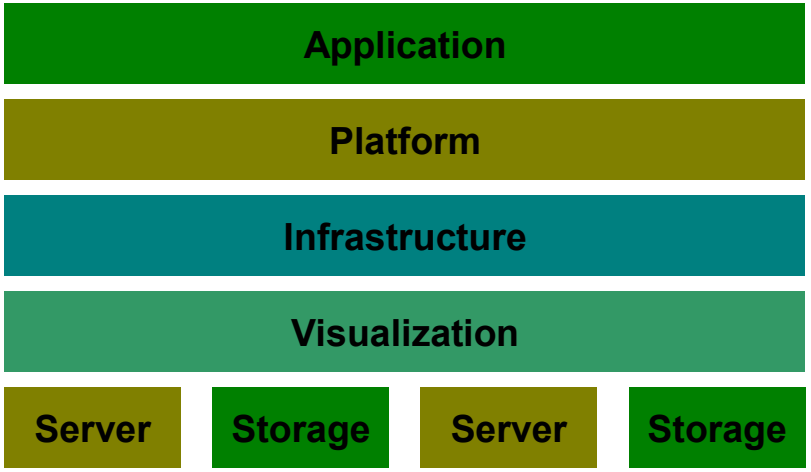


图1-7 云计算的服务模式和类型

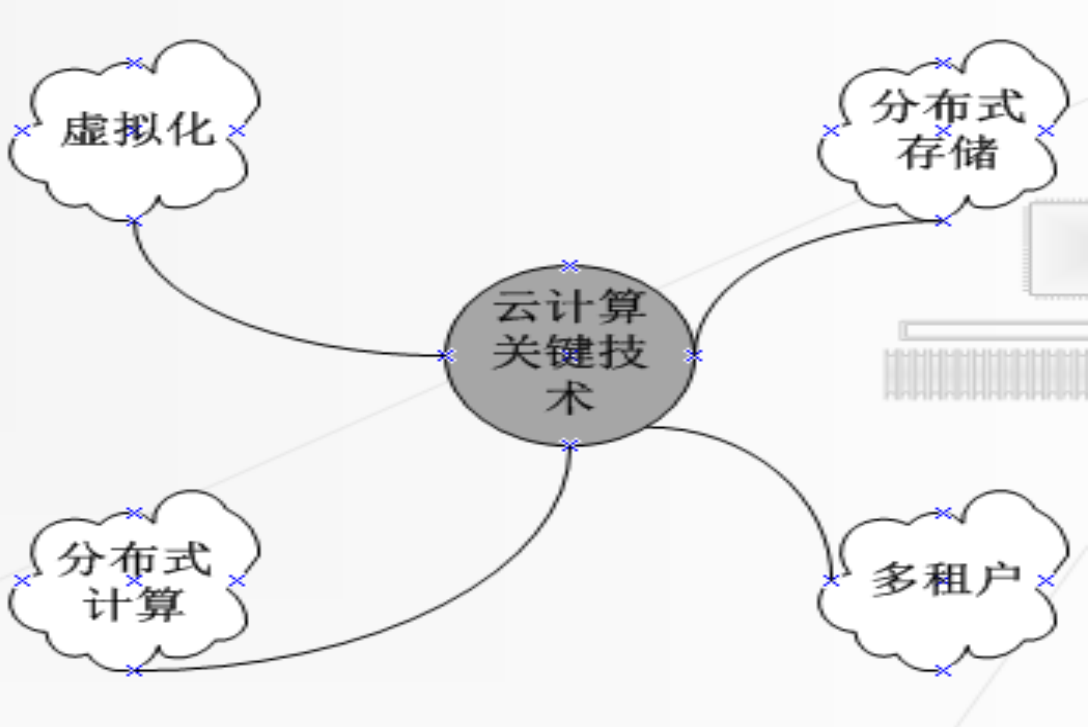
SaaS	从一个集中的系统部署软件，使之在一台本地计算机上(或从云中远程地)运行的一个模型。由于是计量服务，SaaS 允许出租一个应用程序，并计时收费
PaaS	类似于 IaaS，但是它包括操作系统和围绕特定应用的必需的服务
IaaS	将基础设施(计算资源和存储)作为服务出租



SaaS	Software as a Service	Google Apps, Microsoft “Software+Services”
PaaS	Platform as a Service	IBM IT factory, Google App Engine, Force.com
IaaS	Infrastructure as a Service	Amazon EC2, IBM Blue Cloud, Sun Grid

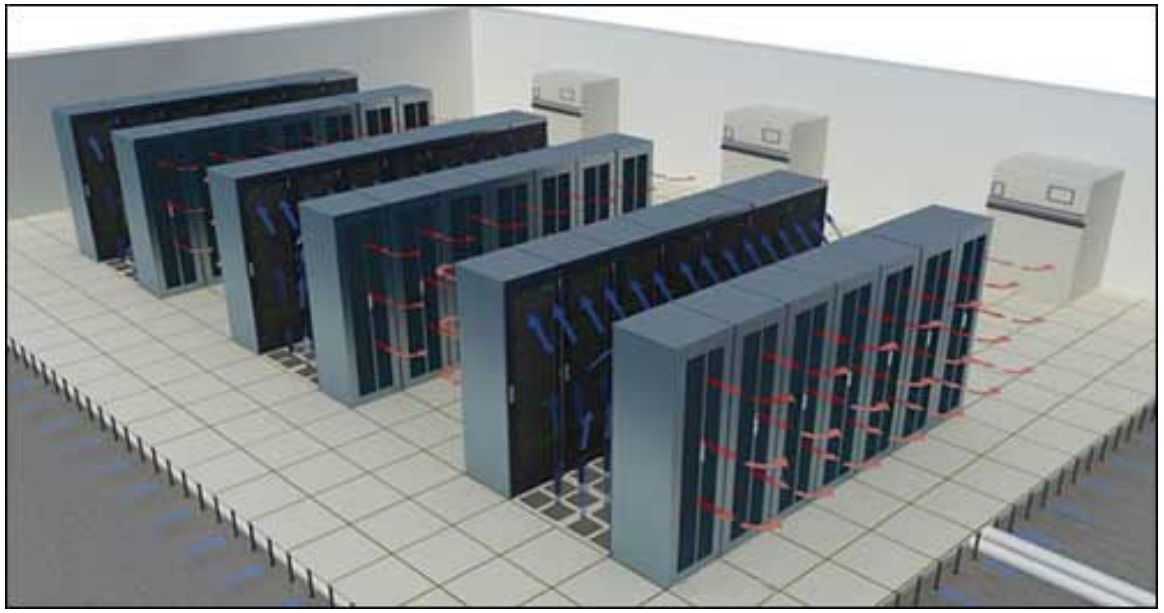
2. 云计算关键技术

- 云计算关键技术包括：虚拟化、分布式存储、分布式计算、多租户等.



3. 云计算数据中心

- 云计算数据中心是一整套复杂的设施，包括刀片服务器、宽带网络连接、环境控制设备、监控设备以及各种安全装置等；
- 数据中心是云计算的重要载体，为云计算提供计算、存储、带宽等各种硬件资源，为各种平台和应用提供运行支撑环境；
- 全国各地推进数据中心建设。



4.云计算应用

- 政务云上可以部署公共安全管理、容灾备份、城市管理、应急管理、智能交通、社会保障等应用，通过集约化建设、管理和运行，可以实现信息资源整合和政务资源共享，推动政务管理创新，加快向服务型政府转型；
- 教育云可以有效整合幼儿教育、中小学教育、高等教育以及继续教育等优质教育资源，逐步实现教育信息共享、教育资源共享及教育资源深度挖掘等目标；
- 中小企业云能够让企业以低廉的成本建立财务、供应链、客户关系等管理应用系统，大大降低企业信息化门槛，迅速提升企业信息化水平，增强企业市场竞争力；
- 医疗云可以推动医院与医院、医院与社区、医院与急救中心、医院与家庭之间的服务共享，并形成一套全新的医疗健康服务系统，从而有效地提高医疗保健的质量……

5.云计算产业

- 云计算产业作为战略性新兴产业，近些年得到了迅速发展，形成了成熟的产业链结构，产业涵盖硬件与设备制造、基础设施运营、软件与解决方案供应商、基础设施即服务（IaaS）、平台即服务（PaaS）、软件即服务（SaaS）、终端设备、云安全、云计算交付/咨询/认证等环节。

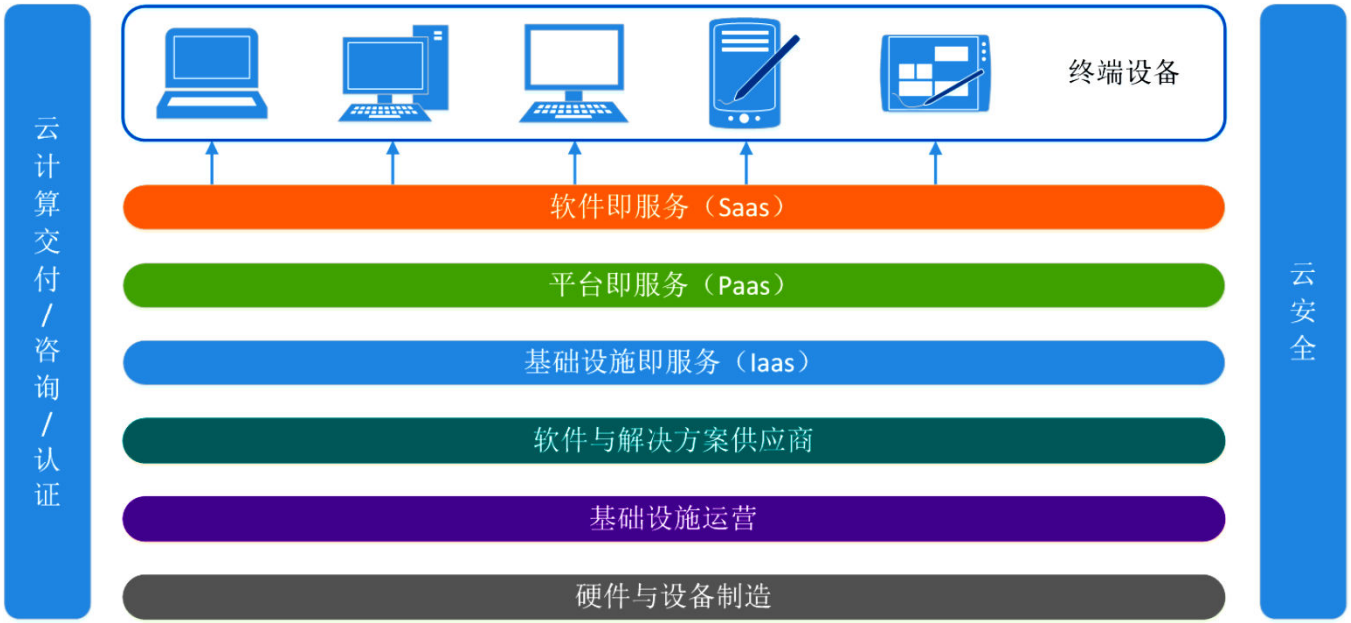


图1-8 云计算产业链

1.8.2 物联网

1. 物联网概念

- 物联网是物物相连的互联网，是互联网的延伸，它利用局部网络或互联网等通信技术把传感器、控制器、机器、人员和物等通过新的方式联在一起，形成人与物、物与物相联，实现信息化和远程管理控制.

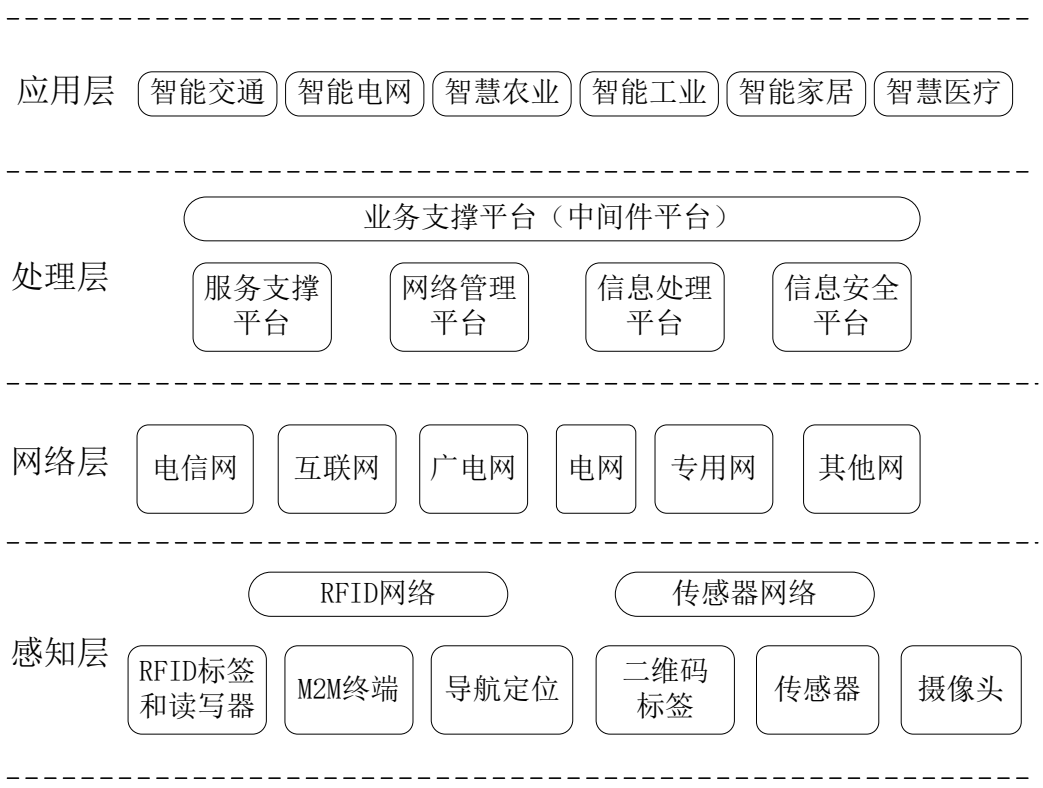


图1-9 物联网体系架构

2. 物联网关键技术

- 物联网中的关键技术包括识别和感知技术（二维码、RFID、传感器等）、网络与通信技术、数据挖掘与融合技术等。

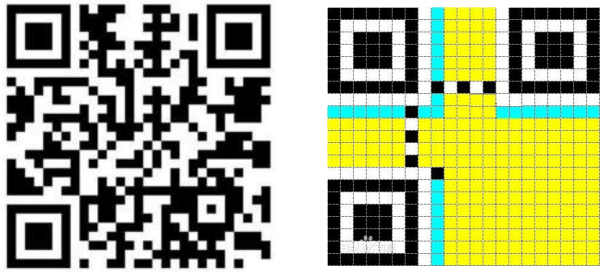


图1-10 矩阵式二维码

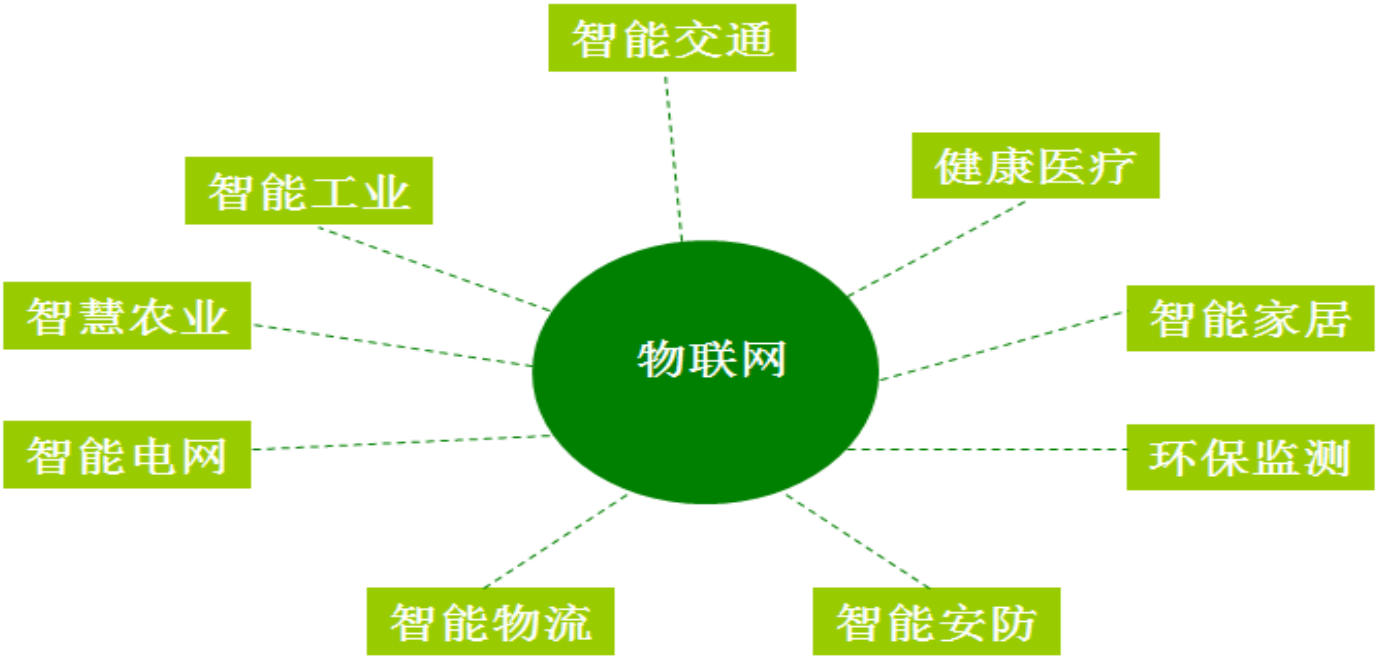
图1-11 采用RFID芯片的公交卡

(a)温湿度传感器　(b)压力传感器　(c)烟雾传感器

图1-12 不同类型的传感器

3.物联网应用

- 物联网已经广泛应用于智能交通、智慧医疗、智能家居、环保监测、智能安防、智能物流、智能电网、智慧农业、智能工业等领域，对国民经济与社会发展起到了重要的推动作用



4.物联网产业

- 完整的物联网产业链主要包括核心感应器件提供商、感知层末端设备提供商、网络提供商、软件与行业解决方案提供商、系统集成商、运营及服务提供商等六大环节。

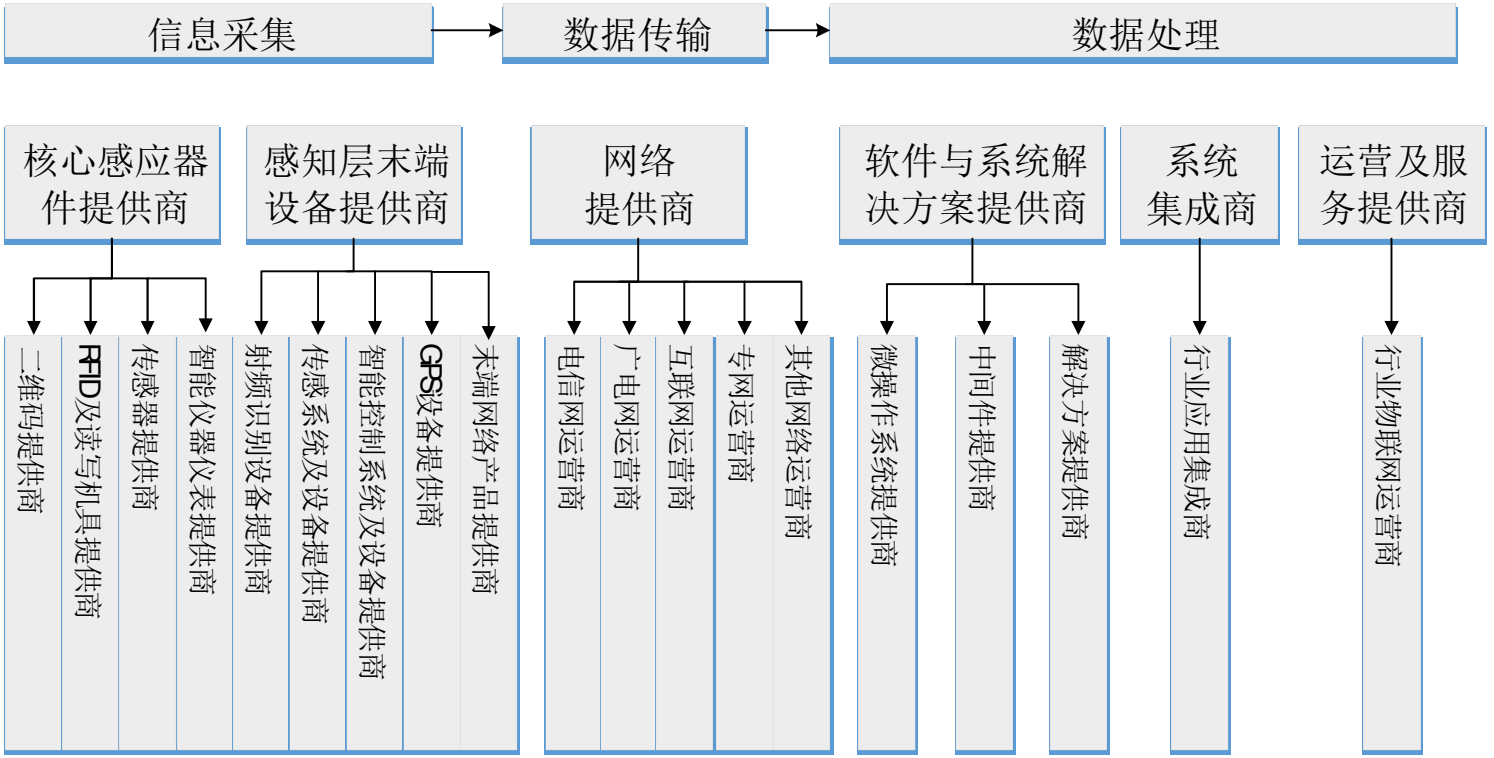


图1-13 物联网产业链

1.8.3大数据与云计算物联网关系

- 云计算、大数据和物联网代表了IT领域最新的技术发展趋势，三者既有区别又有联系。

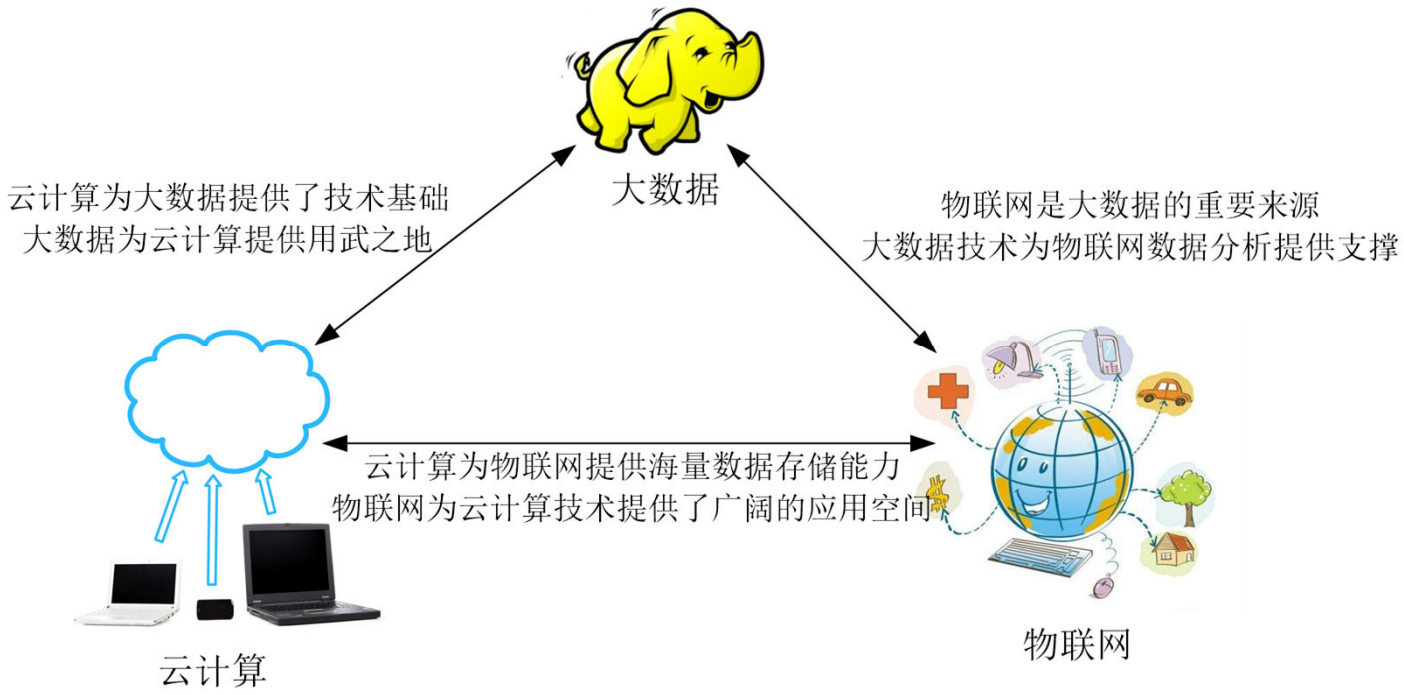


图1-14 大数据、云计算和物联网之间的关系

本章小结

- 本章介绍了大数据技术的发展历程，并指出信息科技的不断进步为大数据时代提供了技术支撑，数据产生方式的变革促成了大数据时代的来临；
- 大数据具有数据量大、数据类型繁多、处理速度快、价值密度低等特点，统称“**4V**”。大数据对科学研究、思维方式、社会发展、就业市场和人才培养等方面，都产生了重要的影响，深刻理解大数据的这些影响，有助于我们更好把握学习和应用大数据的方向；
- 大数据在金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业都得到了日益广泛的应用，深刻地改变着我们的社会生产和日常生活；
- 大数据并非单一的数据或技术，而是数据和大数据技术的综合体。大数据技术主要包括数据采集、数据存储和管理、数据处理与分析、数据安全和隐私保护等几个层面的内容；
- 大数据产业包括**IT**基础设施层、数据源层、数据管理层、数据分析层、数据平台层和数据应用层，在不同层面，都已经形成了一批引领市场的技术和企业；
- 本章最后介绍了云计算和物联网的概念和关键技术，并阐述了大数据、云计算和物联网三者之间的区别与联系。