大数据技术公开课
《大数据概念、技术与应用》
2015年10月13日 山东大学

山东大学公开课主页

第1讲 大数据概述

林子雨 博士/助理教授

厦门大学计算机科学系

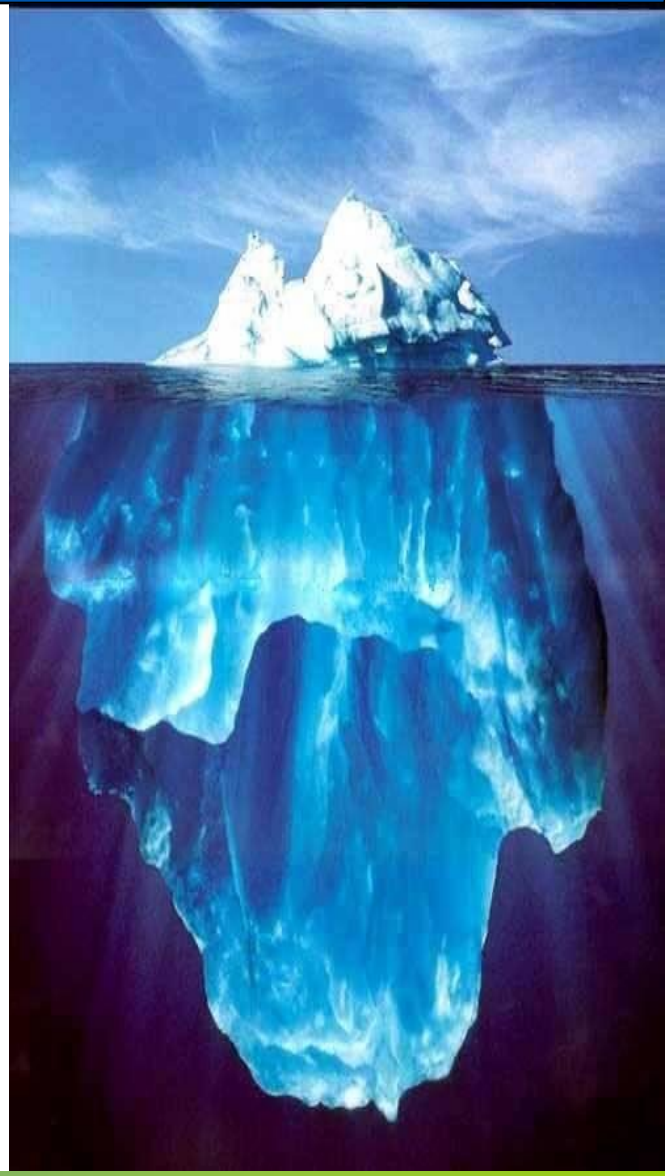
厦门大学云计算与大数据研究中心

E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

提纲

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 大数据与云计算、物联网的关系
1.8产业化应用案例分享



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
SOFTWARE
ANALYTICS
SIZE
NEEDED
PETABYTES
INTERNET
TECHNOLOGIES
MANAGEMENT
SOCIAL
LARGE
EVERY
STORAGE
COMPLEX
DATABASES
RESEARCH
INFORMATION
NETWORKS
INCLUDE
SYSTEMS
SHARED
TIME
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

- 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

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

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

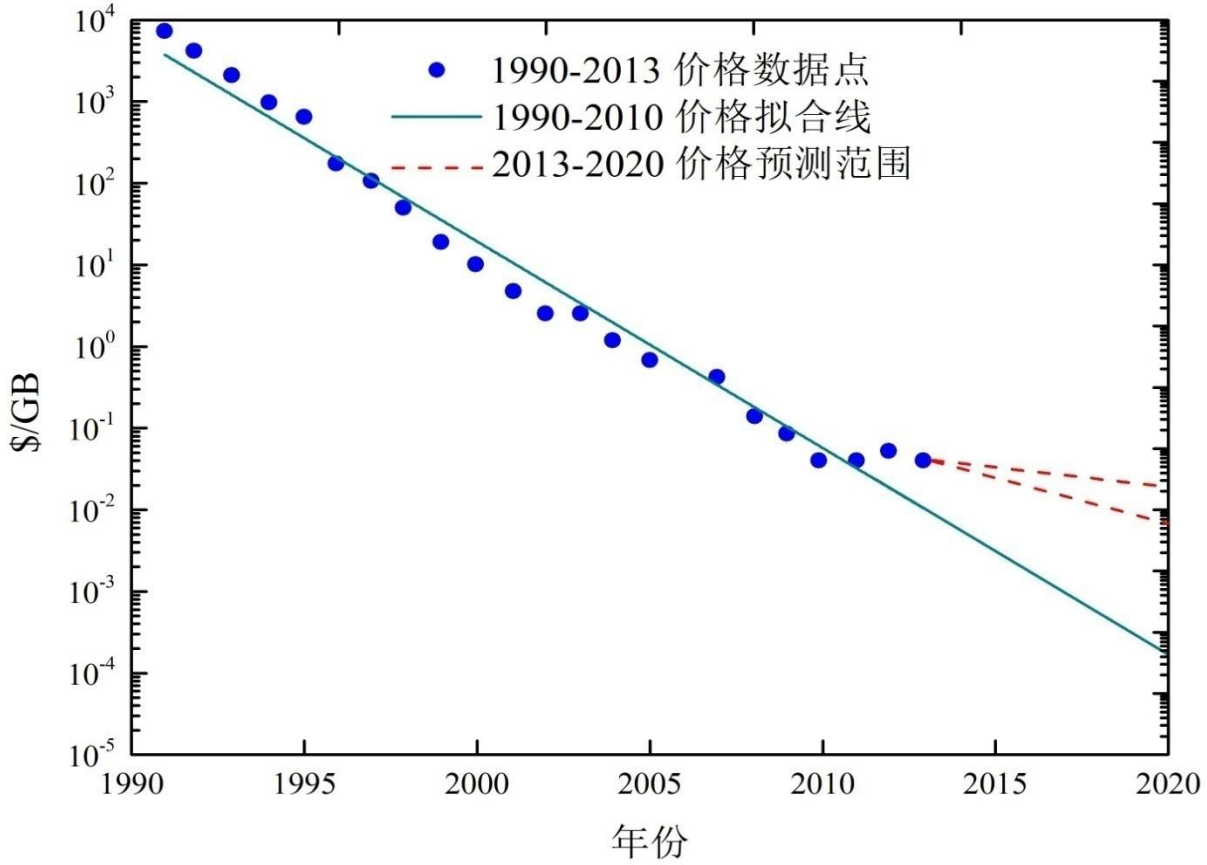


图1-1 存储价格随时间变化情况

1.1.2信息科技为大数据时代提供技术支撑

2. CPU处理能力大幅提升

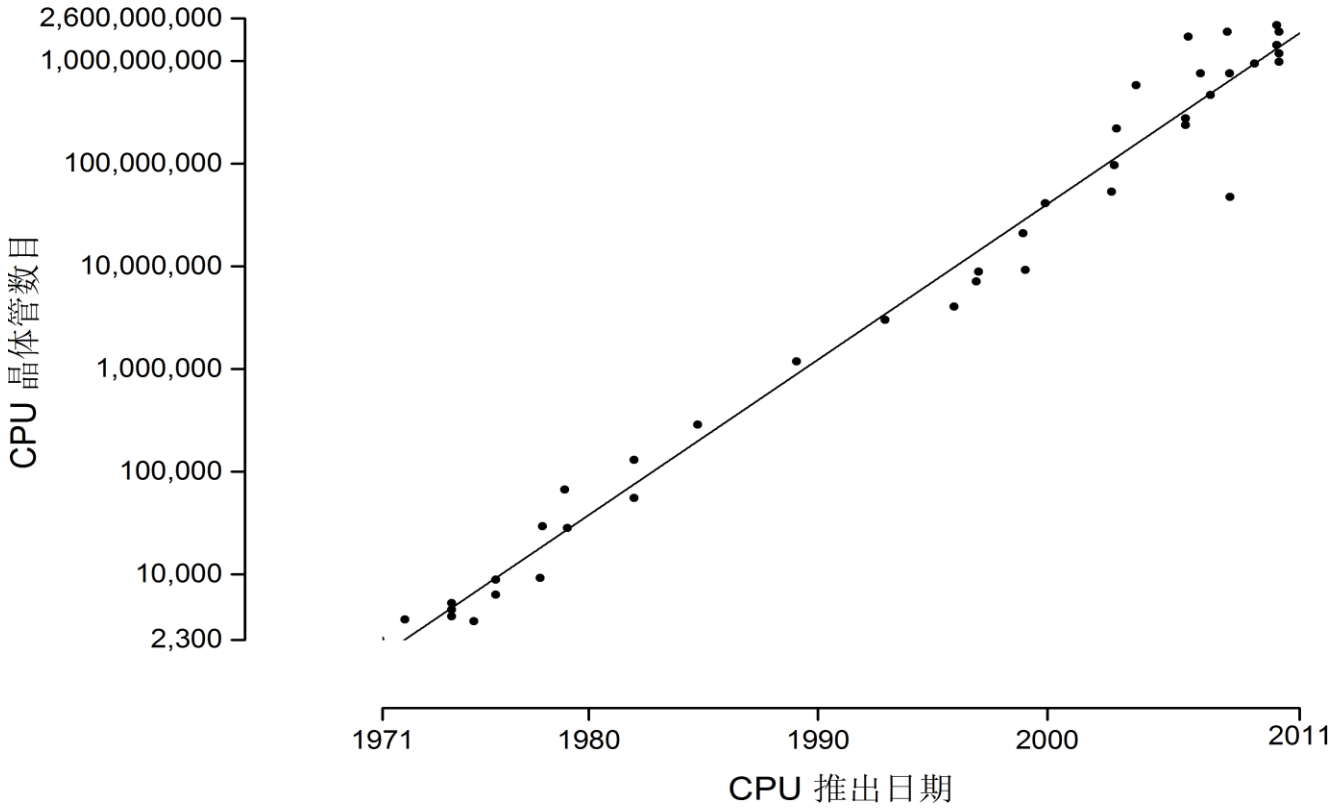


图1-3 CPU晶体管数目随时间变化情况

1.1.2信息科技为大数据时代提供技术支撑

3. 网络带宽不断增加

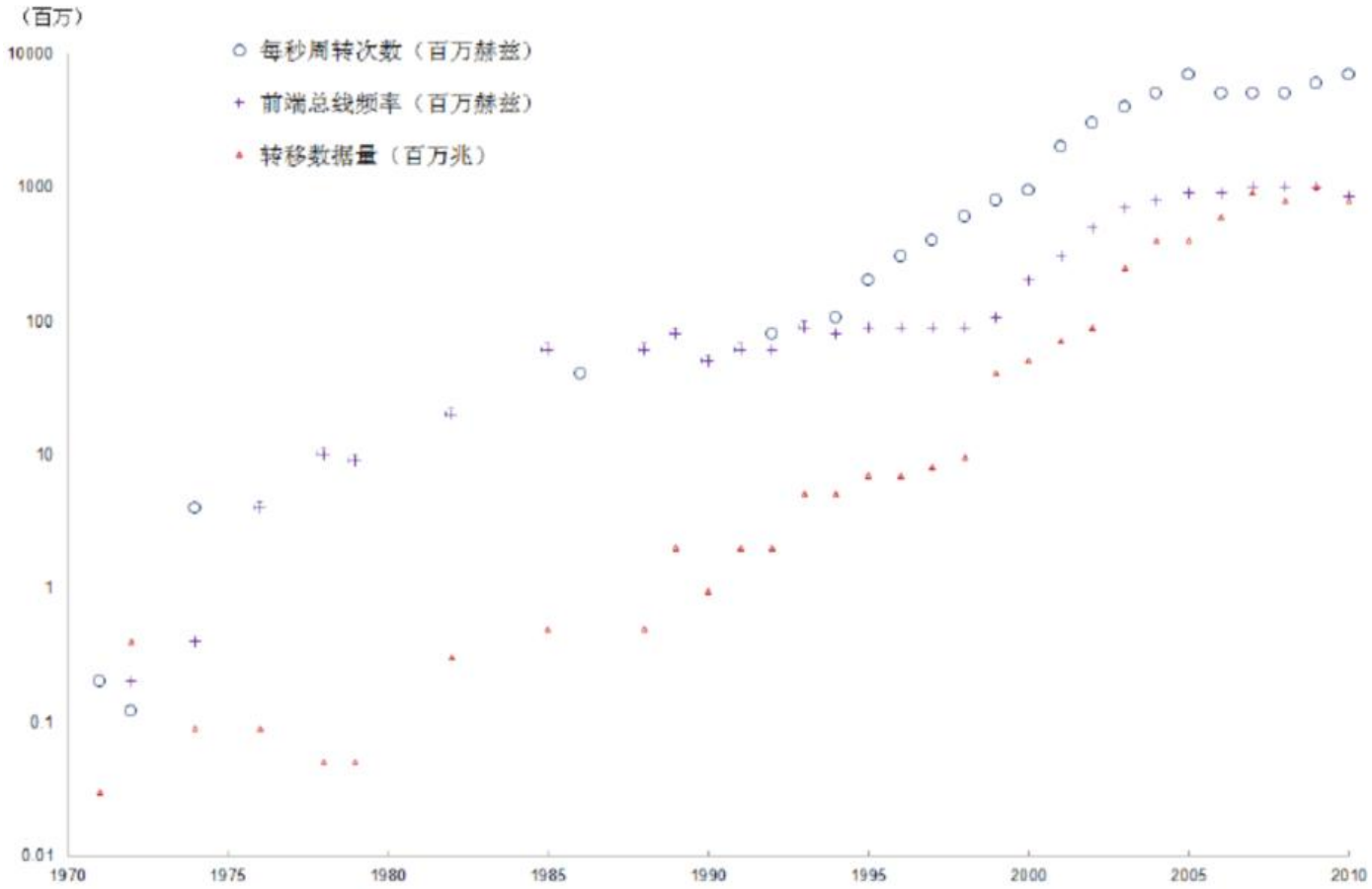


图1-4 网络带宽随时间变化情况

1.1.3数据产生方式的变革促成大数据时代的来临

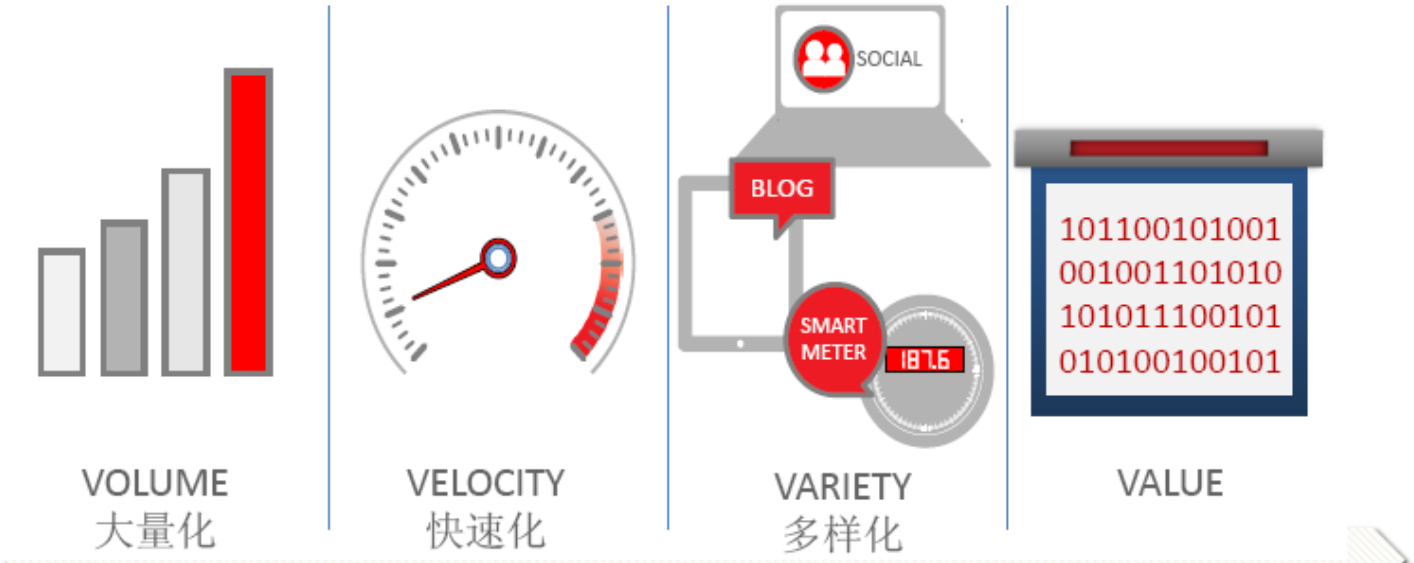
图1-5 数据产生方式的变革

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GFD和MapReduce等发数据技术受到追捧，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

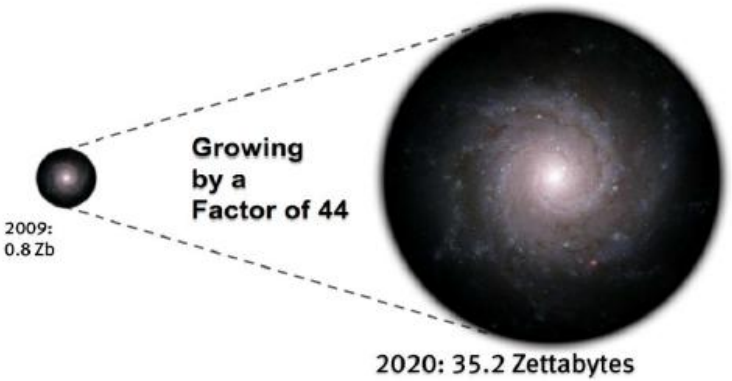
1.2大数据概念

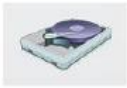


大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

1.2.1数据量大

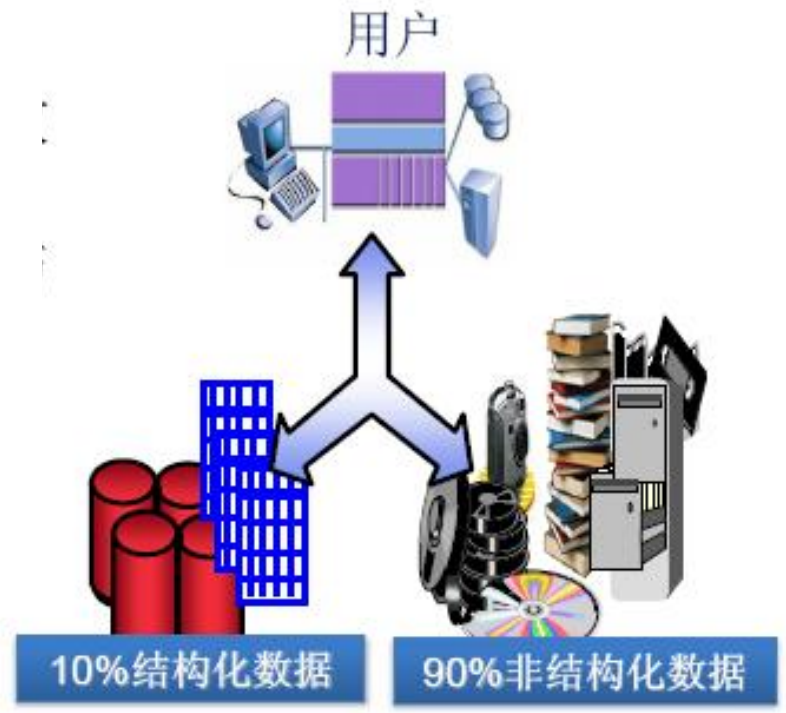
- 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）
- 人类在最近两年产生的数据量相当于之前产生的全部数据量
- 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍



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

1.2.2数据类型繁多

- 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
- 非结构化数据类型多样
  - 邮件、视频、微博
  - 位置信息、链接信息
  - 手机呼叫、网页点击
  - “长微博”



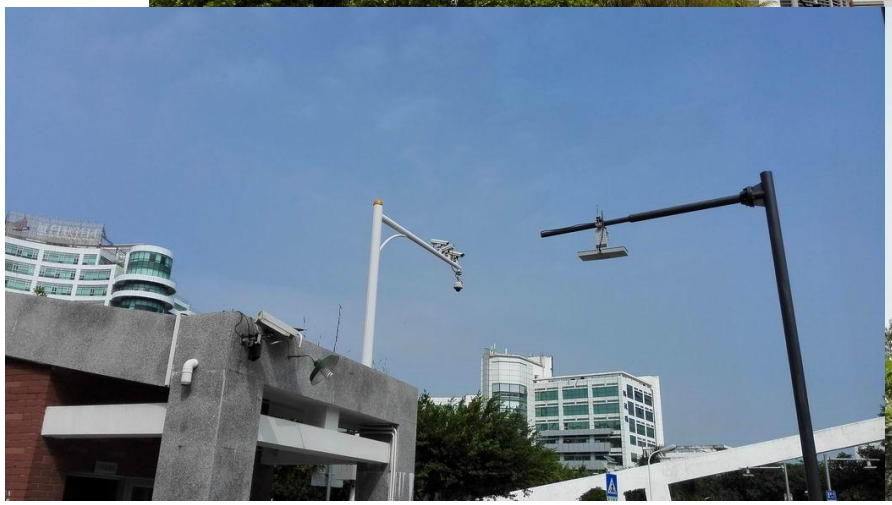
1.2.3处理速度快

- 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

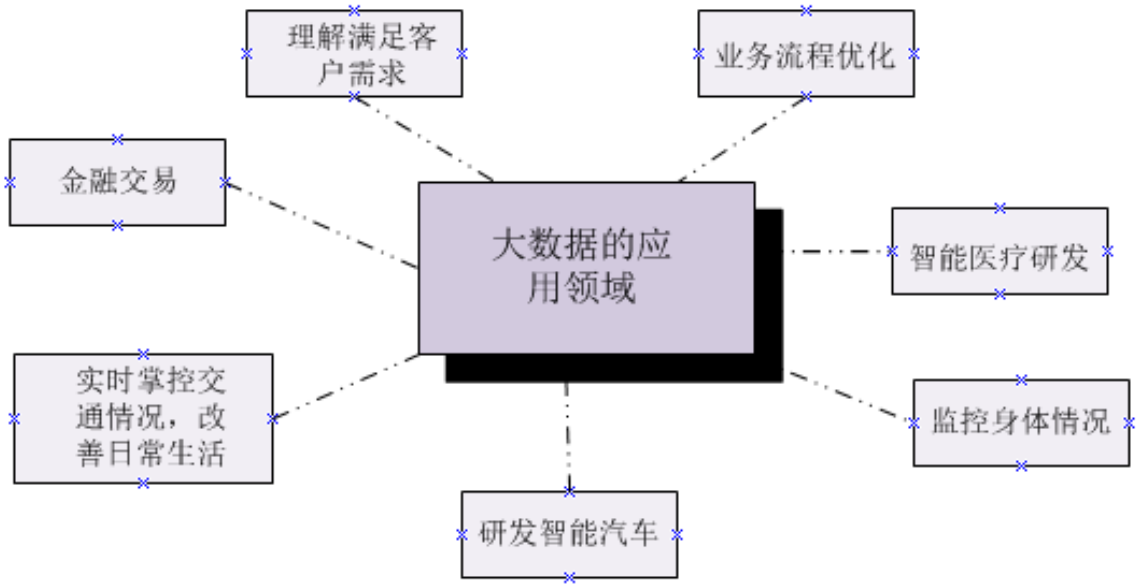


1.3大数据的影响

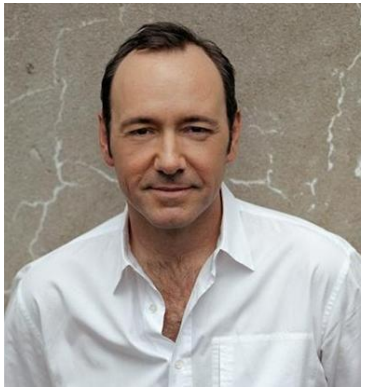
- 大数据对科学研究、思维方式和社会发展都具有重要而深远的影响。
- 在科学研究方面，大数据使得人类科学研究在经历了实验、理论、计算三种范式之后，迎来了第四种范式——数据
- 在思维方式方面，大数据具有“全样而非抽样、效率而非精确、相关而非因果”等三大显著特征，完全颠覆了传统的思维方式
- 在社会发展方面，大数据决策逐渐成为一种新的决策方式，大数据应用有力促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现
- 在就业市场方面，大数据的兴起使得数据科学家成为热门职业
- 在人才培养方面，大数据的兴起，将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制

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

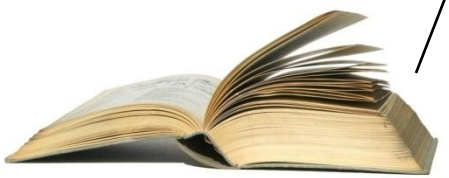


典型的大数据应用实例

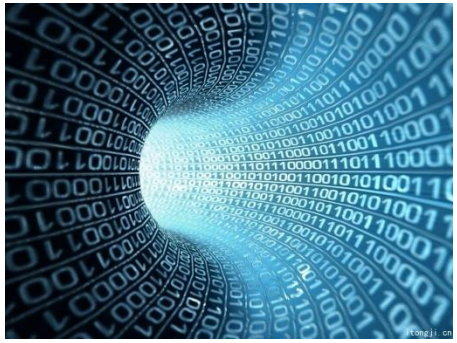


Kevin Spacey

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》

典型的大数据应用实例

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1-5 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

1.7大数据计算模式

表1-3 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大数据与云计算、物联网的关系

- 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，三者相辅相成，既有联系又有区别

1.7.1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

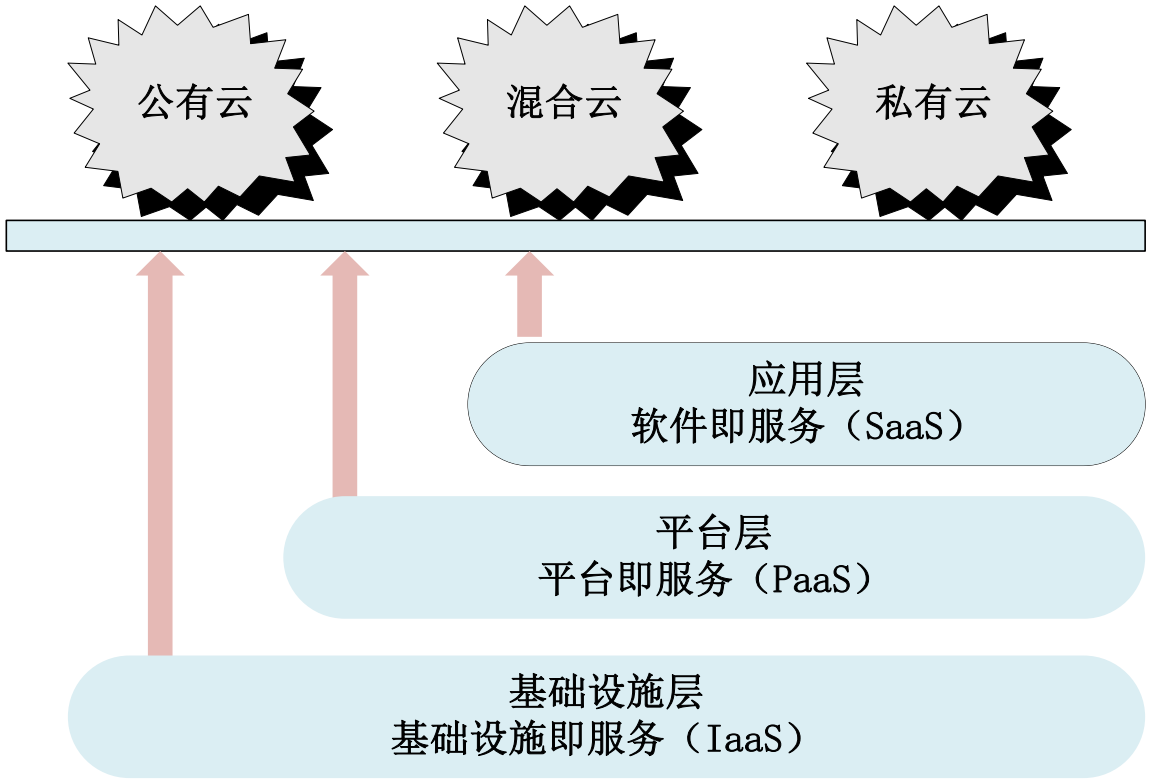
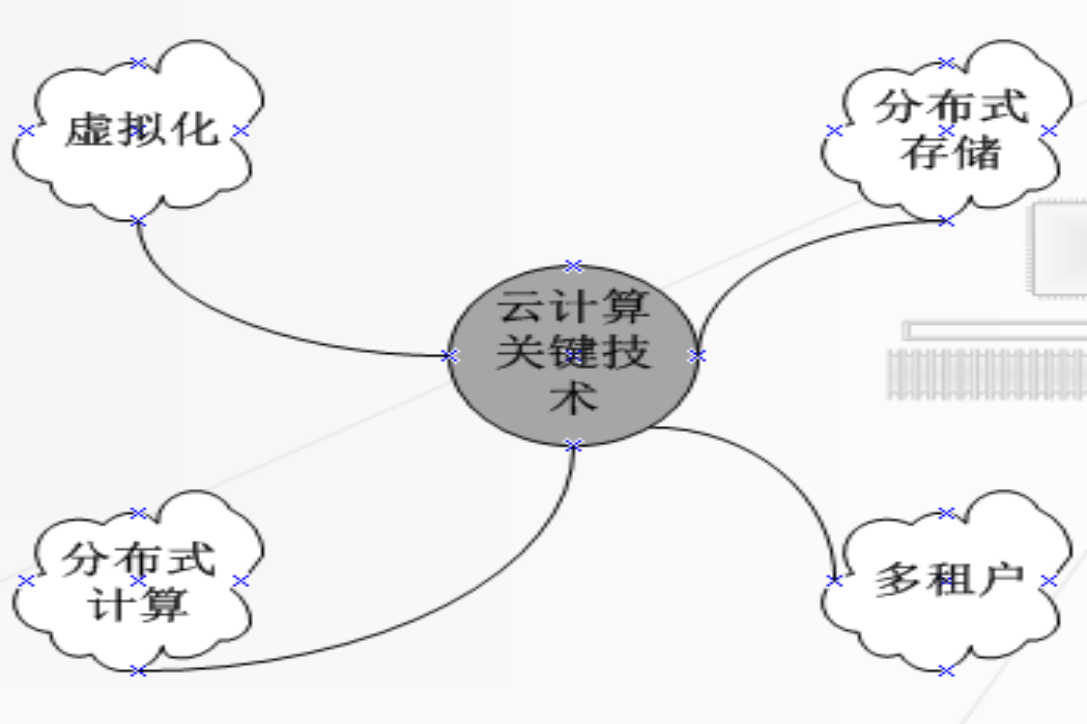


图1-7 云计算的服务模式和类型

1.7.1云计算

2. 云计算关键技术

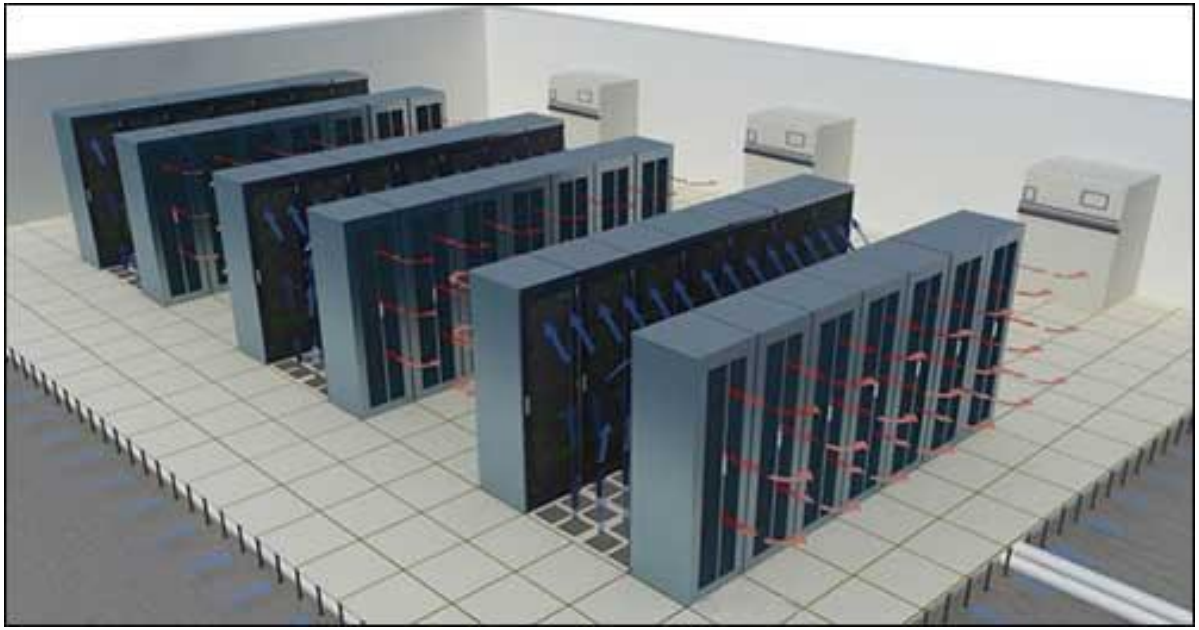
- 云计算关键技术包括：虚拟化、分布式存储、分布式计算、多租户等



1.7.1云计算

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



1.8.1云计算

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

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

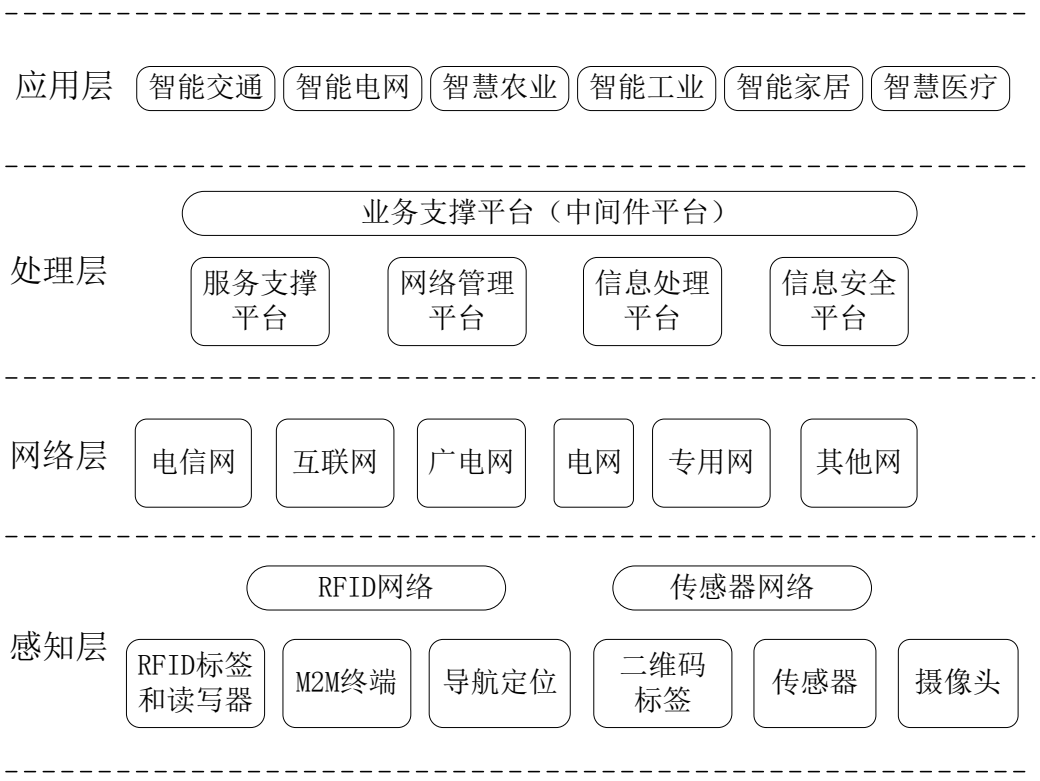


图1-9 物联网体系架构

1.8.2物联网

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

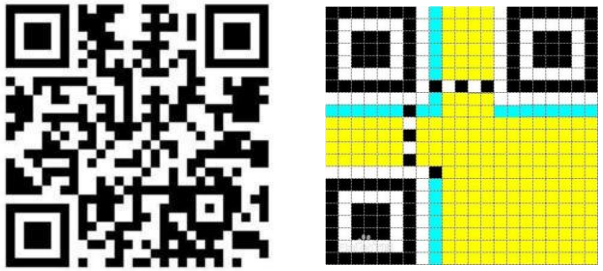


图1-10 矩阵式二维码

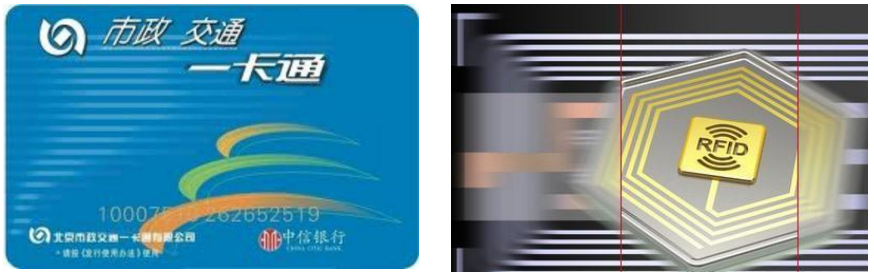


图1-11 采用RFID芯片的公交卡

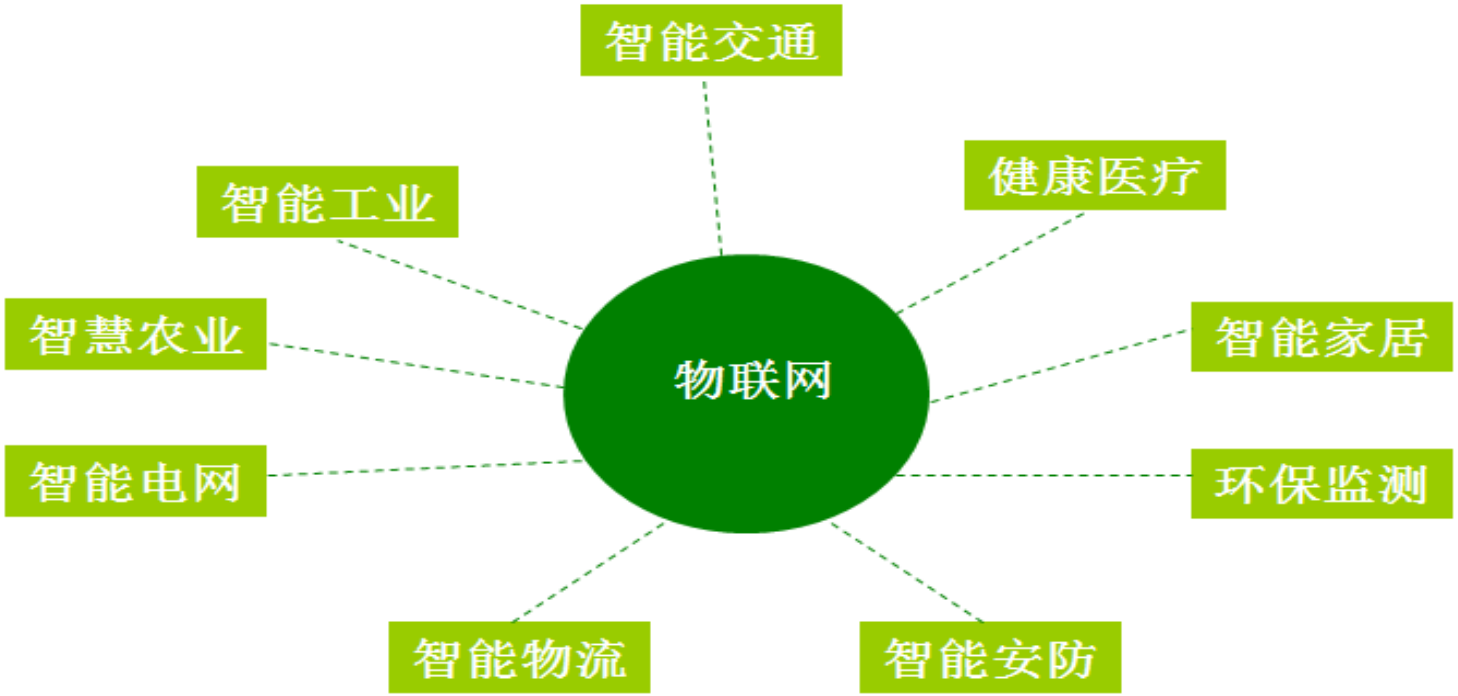
(a)温湿度传感器　(b)压力传感器　(c)烟雾传感器

图1-12 不同类型的传感器

1.8.2物联网

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



1.8.3大数据与云计算、物联网的关系

- 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

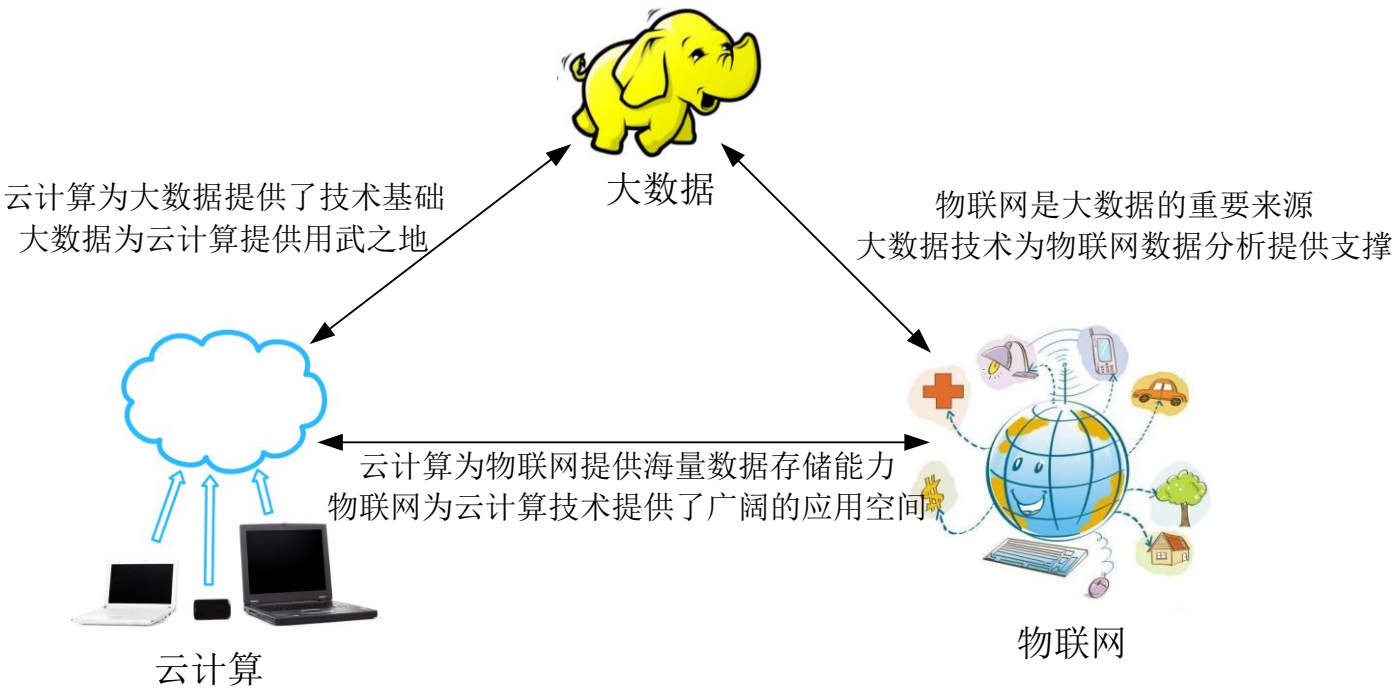


图1-9 大数据、云计算和物联网之间的关系

1.8产业化应用案例分享

案例一：在物流行业的应用

物流信息化领域相关工作

2014年9月，主持《石狮市物流园区建设方案研究》课题

2014年6月，泉州市物流信息化产业技术创新战略联盟”成立

2015年4月、8月给厦门物流协会专题讲座

物流行业应用案例：智能物流

智能物流集成商案例：阿里巴巴的中国智能物流骨干网（地网）

菜鸟网络到底是什么?

- 中国智能物流骨干网，又名“菜鸟”
- **菜鸟网络计划在5到8年内，打造一个全国性的超级物流网。**
- 这个网络能在24小时内将货物运抵国内任何地区，能支撑日均300亿元(年度约10万亿元)的巨量网络零售额。

1000亿元投资物流基础设施　强强联手共建智能骨干网络物流信息系统向所有的制造商、网商、快递公司、第三方物流公司完全开放

阿里物流体系

天网	地网
天猫牵头负责与各大物流快递公司对接的数据平台	即“菜鸟”，又称“中国智能物流骨干网（CSN）”

中国智能物流骨干网——菜鸟网络

依托阿里巴巴集团旗下多个电商平台为核心的大数据平台（**天网**），即掌握的网络购物物流需求数据、电商货源数据、货流量及分布数据、以及消费者长期购买习惯数据，优化仓储选址、干线物流基础设施建设、以及物流体系建设

关键举措一：智能化建立物流集散中心（基础设施平台），搭建骨干网框架

建立统一的仓储及调度体系，整合和集中管理原本各快递公司自建的物流体系

关键举措二：整合所有服务商信息系统，实现骨干网内部信息统一

关键举措三：应用智能化技术，补足物流行业仓储环节短板

采用自动分拣、自动传输、自动出库、自动补货等手段建立智能实体仓库，在减少库存积压的基础上提升效率，同时建立虚拟仓库，实现信息与数据对接的信息化管理

关键举措四：构建开放数据应用平台，向物流生态系统内各种群提供服务

构建向“电子商务企业、物流公司、仓储企业、第三方物流服务商以及供应链服务商”开放的数据应用平台

产业化应用案例分享

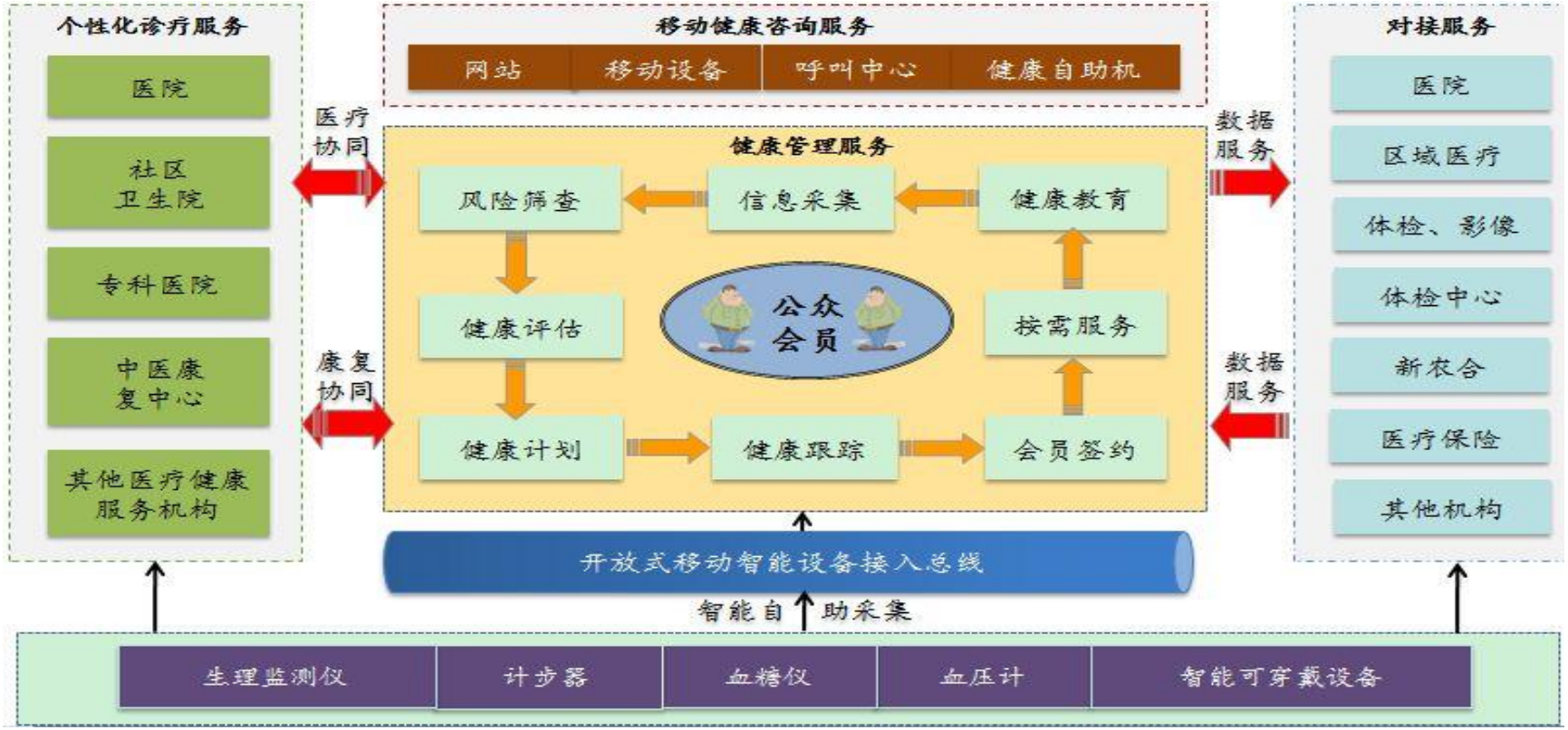
案例二：在医疗健康行业的应用

基于大数据的综合健康服务平台

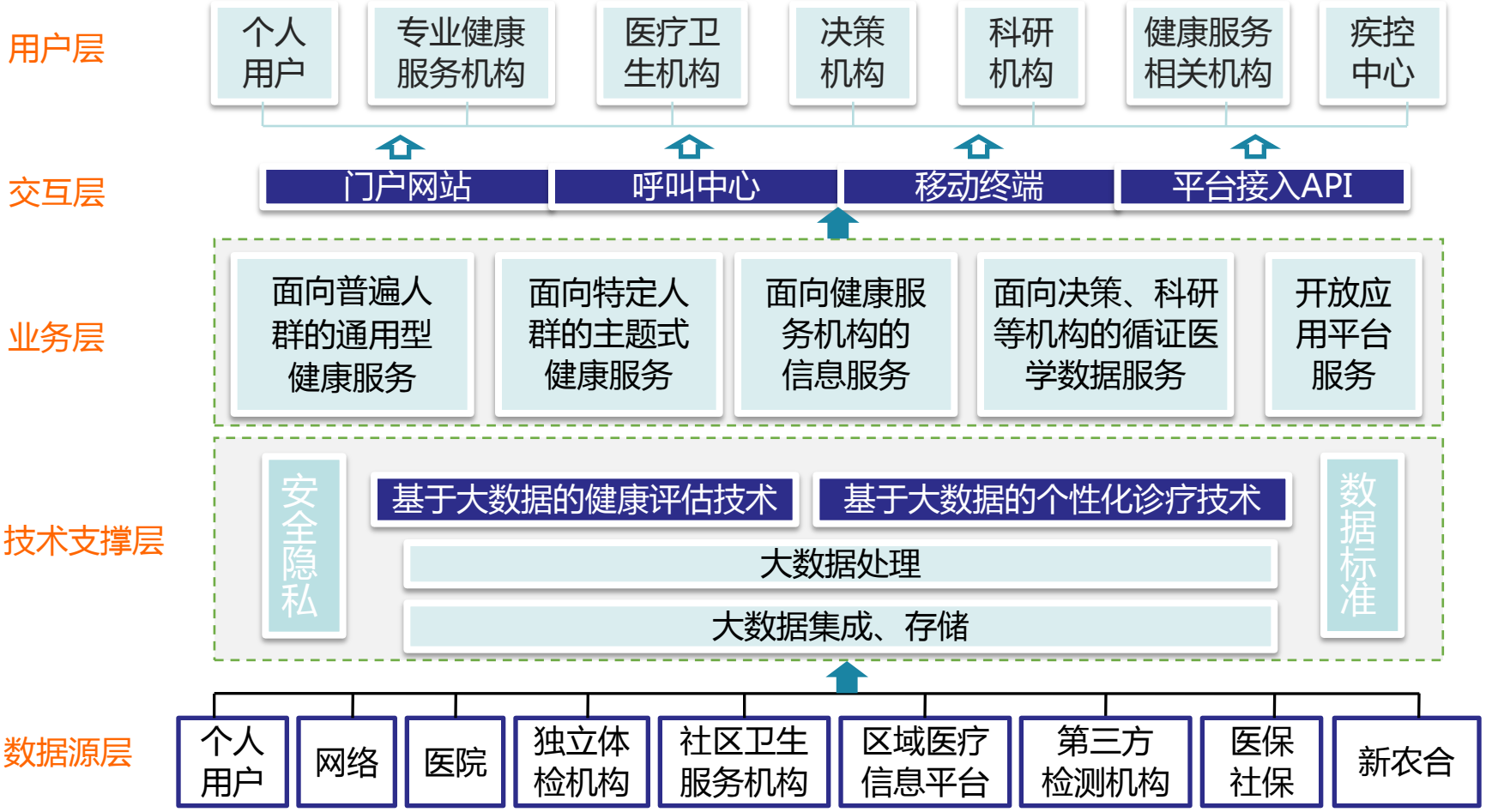
2014年，由四个合作单位组建的团队
封闭写作平台项目申请书

医疗健康行业应用：综合健康服务平台

建设目标：构建覆盖全生命周期、内涵丰富、结构合理的以人为本全面连续的综合健康服务体系，利用大数据技术和智能设备技术，提供线上线下相结合的公众健康服务，实现“未病先防、已病早治、既病防变、愈后防复”，满足社会公众多层次、多方位的健康服务需求，提升人民群众的身心健康水平。



医疗健康行业应用：综合健康服务平台



产业化应用案例分享

案例三：在餐饮配送行业的应用

2015年1月9日，在泉州举行云配送产品推介会

云配送

“云配送”系统是一款基于云计算技术的在线软件系统，以微信平台系统为技术支撑，微信用户为目标消费群体，服务于全国线下实体商家及线下网络商家的微信营销系统产品。

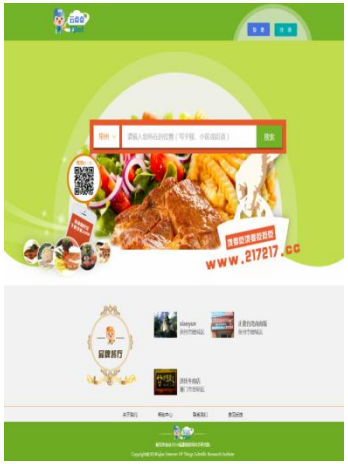
1. 订单形成
2. 进度查询
3. 在线支付
4. 数据分析

……

云配送产品特点

- 产品特点
  - 微信下单、手机APP下单、网站下单
  - 商家打印机打印订单
  - 订单统计分析
- 产品竞争优势
  - 具有多种支付方式
  - 抢单配送
  - 银联POS机与打印机相结合
  - 条形码配送

手机APP下单



网页下单

抢单配送

无线打印机

融合订单打印功能的POS机

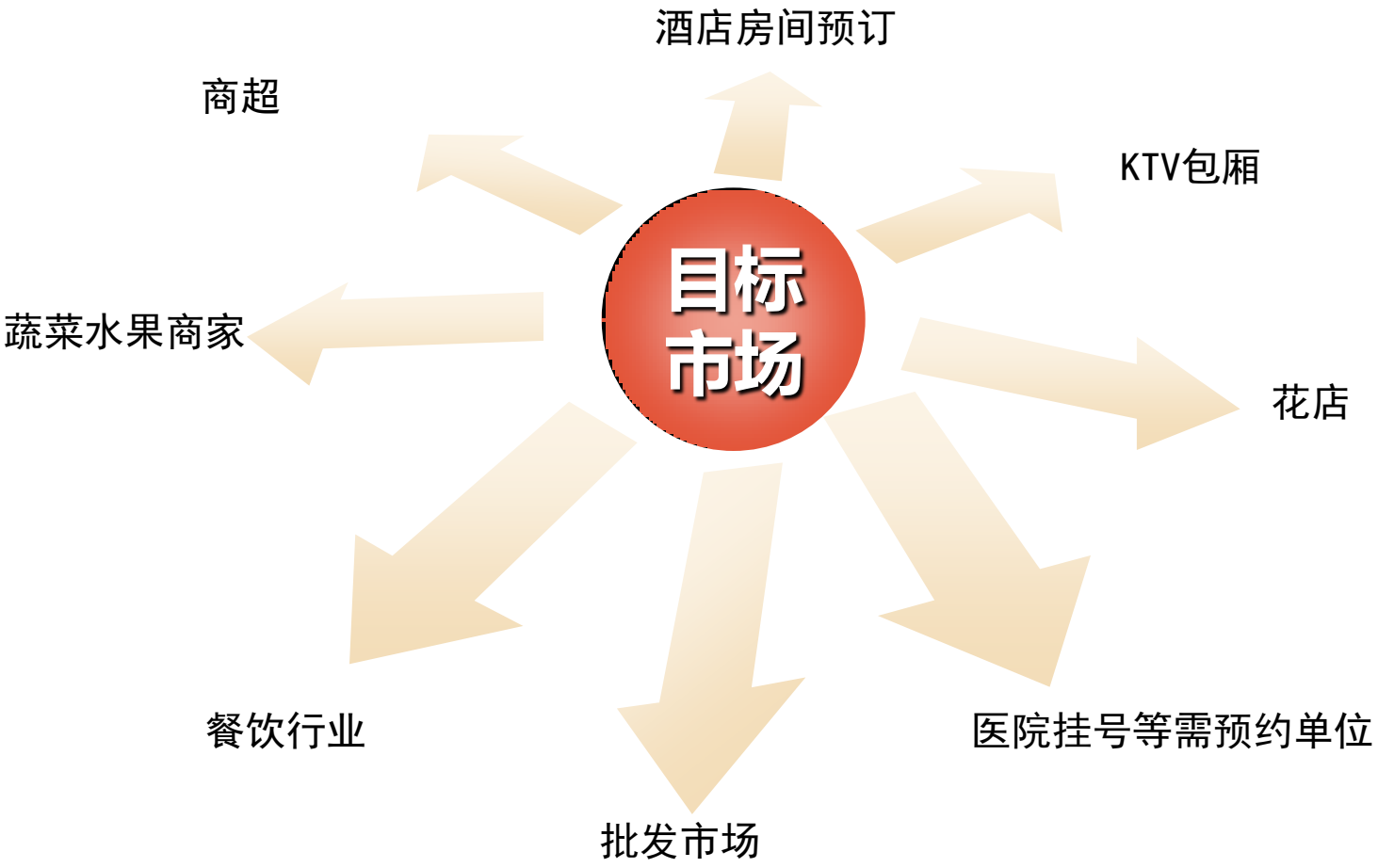
微信下单

云配送产品使用方法

通过关注商家店铺微信二维码，在线下单后，无线打印机立即打印出客户所需的服务

云配送系统目标市场

云配送系统的目标市场



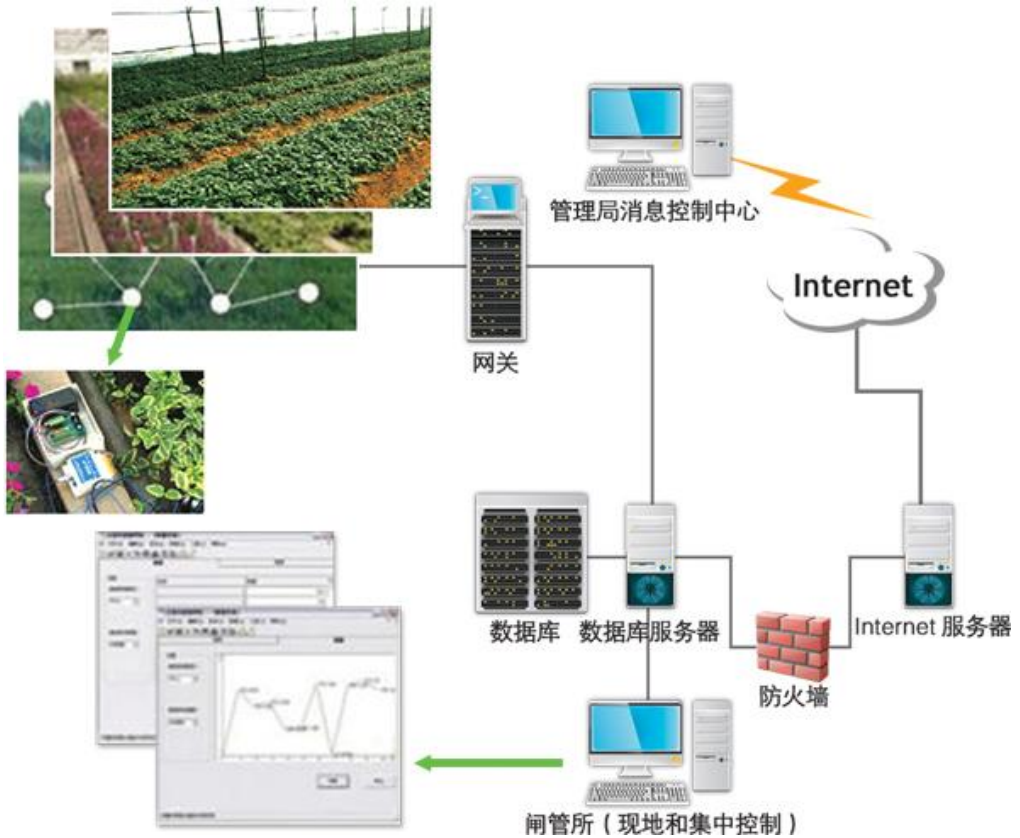
产业化应用案例分享

案例四：在菜篮子工程中的应用

物联网改变传统农业生产方式

智慧农业

智慧农业是农业生产的高级阶段，是集新兴的互联网、移动互联网、云计算和物联网技术为一体，依托部署在农业生产现场的各种传感节点（环境温湿度、土壤水分、二氧化碳、图像等）和无线通信网络实现农业生产环境的智能感知、智能预警、智能决策、智能分析、专家在线指导，为农业生产提供精准化种植、可视化管理、智能化决策。



物联网改变传统农业生产方式

2014年，调研福建南安绿莹生态农业基地

智慧农业

本讲小结

- 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

主讲教师

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度厦门大学奖教金获得者。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，编著出版中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》并成为畅销书籍，编著并免费网络发布40余万字中国高校第一本闪存数据库研究专著《闪存数据库概念与技术》；主讲厦门大学计算机系本科生课程《数据库系统原理》和研究生课程《分布式数据库》《大数据技术基础》。具有丰富的政府和企业信息化培训经验，曾先后给中国移动通信集团公司、福州马尾区政府、福建省物联网科学研究院、石狮市物流协会、厦门市物流协会等多家单位和企业开展信息化培训，累计培训人数达2000人以上。

大数据学习教材推荐

《大数据技术原理与应用——概念、存储、处理、分析与应用》，由厦门大学计算机科学系林子雨博士编著，是中国高校第一本系统介绍大数据知识的专业教材。

全书共有13章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网

Department of Computer Science, Xiamen University, October, 2015