《大数据技术》

第1章 大数据概述

洪韬

E-mail: 2993400893@qq.com

课程安排

- **课程构成：**

  理论课+实践课

  理论课着重在于大数据整体**知识体系架构**、**大数据处理流程**以及当前主流大数据组件的设计思想、运行原理；

  实验课着重在于学习掌握搭建大数据平台，对大数据平台中重要组件编程实践进行练习，能够学会使用某些重要组件比如MR编程、Hbase、HiveQL等。

- **章节顺序：**

  第一部分：大数据概述

  第一章、第二章；

  第二部分：大数据存储与管理

  第三到第五章；

  第三部分：大数据处理分析

  第七到第十三章；

  第四部分：大数据应用实例分析

  第十五章，以推荐系统为例，讲解大数据综合应用的整个流程；

课程安排

- **课程目标：**

了解掌握**大数据体系架构**，掌握大数据平台下各重要模块的**设计思路**、**运行原理**，有一定的编**程实践处理简单问题的能力**。学习理解大数据处理问题的思维过程，并能够将大数据的一些思维、解决问题方法运用或者结合到日后的工作中。（比如，可以爬取驻马店楼盘数据，使用工具分析出驻马店房地产行业特点-地理位置和房价等）

- **课程考核：**

本课程为专业拓展课，为考查课，平时成绩、实践成绩和期末考核比例为2：3：5。课堂中会有一些提问算到平时成绩中。

考核环节	比例	考核/评价细则
平时考核	20	**日常考勤（50%）+上课问答（50%）**
实验（实训）考核	30	**实验考勤（30%）+实验报告（70%）**
期末考核	50	**理论课各章节知识点综合试卷考察/基于一个数据处理应用的大作业考察**

大数据概述

1.1 什么是大数据

1.2 大数据时代

1.3 大数据概念

1.4 大数据的影响

1.5 大数据的应用

1.6 大数据关键技术

1.7 大数据计算模式

1.8 大数据产业

1.9 大数据与云计算、物联网的关系

1.1什么是大数据

究竟什么是“大数据”，它又会给我们的生活带来什么影响呢？不如我们用一个在网上流传的段子来解释一下，可能会比较容易理解。

- 一家快餐披萨店，外卖电话响了，店长拿起电话。
- **店长：**您好，这里是××披萨店。请问有什么需要我为您服务？
- **顾客：**你好，我想要订一份披萨。
- **店长：**请问您是陈先生吗？
- **顾客：**你怎么知道我姓陈？
- **店长：**陈先生，因为我们的CRM（客户关系管理）系统对接了三大通讯服务商，看到您的来电号码，我就知道您贵姓了。
- **顾客：**哦，那我想要一份海鲜至尊披萨。
- **店长：**陈先生，海鲜披萨不适合您，建议您另选一份。
- **顾客：**为什么？

1.1什么是大数据

- **店长：**根据您的医疗记录，您的血尿酸值偏高，有痛风的症状，建议您不要食用高嘌呤的食物。您可以试试我们店最经典的田园蔬菜披萨，低脂、健康，符合您現阶段的饮食要求。
- **顾客：**你怎么知道我会喜欢这种披萨？
- **店长：**您上周在一家网上书店买了一本《低脂健康食谱》，其中就有这款披萨的菜谱。
- **顾客：**那好吧。我要一个家庭特大号披萨，多少钱？

1.1什么是大数据

- **店长：**99元。这个足够您一家六口吃了。但您的母亲应该少吃，她上个月刚做了心脏搭桥手术，还处于恢复期。
- **顾客：**好的，知道了。我可以刷卡吗？
- **店长：**抱歉，陈先生。请您付现吧，因为您的信用卡已经刷爆了，您现在还欠银行5000元，而且还不包括住房贷款利息。
- **顾客：**那我先去附近的提款机取现金。
- **店长：**陈先生，根据银行记录，您今天已经超过了日提款限额。
- **顾客：**算了，那你们直接把披萨送到我家里吧，家里有现金，你们多久能送到？

1.1什么是大数据

- **店长：**大约30分钟。如果您不想等，可以自己来取。
- **顾客：**为什么？
- 店长：我这边看到您家的地址是某某路东段22号，距离我们店开车只有5分钟路程，您名下登记有一辆车号为×××××××的轿车，这辆车目前正在距离您家不到两分钟车程的地方。如果您等不及，可以回家拿了现金就开车来店里取，这大概要花您10分钟的时间，正好是一个披萨出炉的时间。这样，您总共只需花15至20分钟就可以将披萨拿回家，比我们送货上门要快。
- 顾客差点儿晕倒。

1.1什么是大数据

小说三体中描述的超信息化时代

--社会已经不需要手机、电脑、Ipad了。只需点击身边的物体，像衣柜、墙壁等，随时弹出一个**屏幕，可查阅资料，可上网，可看视频。**

周围每一样物品，像室内的冰箱、电饭煲，室外的树木、栏杆，每一样都**自带传感器**，都是一个“电脑终端”，可**自动联网**，可根据**基因识别**每一个人类。

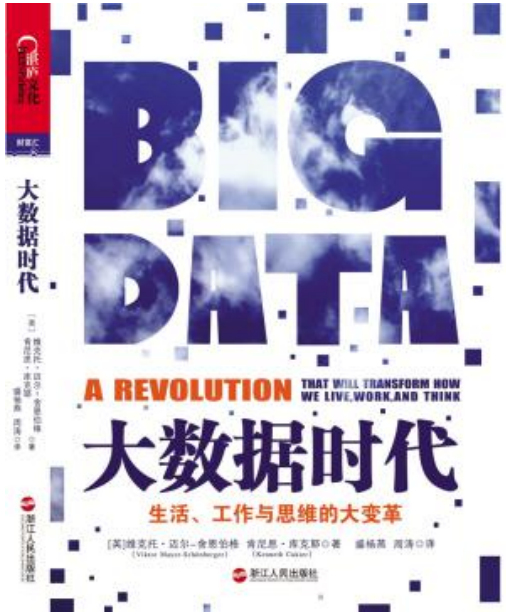
第三次信息化浪潮

- 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

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	**移动互联网**，物联网、云计算和大数据	信息爆炸	Facebook、亚马逊、美团、**字节跳动**、滴滴

数据的爆炸性增长为世界新的时代带来了契机。

1.2大数据时代



我们生活在数据的时代

大家在生活中有没有体会呢？

我们生活在数据的时代-随处皆数据

啤酒与尿布

20世纪90年代沃尔玛超市管理员分析销售数据时，发现了一个很奇怪的现象“啤酒和尿布”两样毫无相关性的数据会经常出现在同一各购物车里。美国学者通过Aprior算法（分析购物篮中的商品集合，找出商品之间关联关系的关联算法，并根据商品之间的关系，找出客户的购买行为）分析了真实情况。

了解到新生儿母亲在家照看孩子，就让年轻的父亲去超市购买孩子需要的尿布。这些父亲会选择既有尿布又有自己喜欢的啤酒的超市，从而能提高买东西的效率。所以说这两样商品齐全的超市会更受年轻父亲喜爱。更有早发现该规律的超市，将啤酒和尿布摆在相邻位置，更加节省了年轻父亲买东西的时间，也就间接提升了超市营业额。

我们生活在数据的时代-随处皆数据

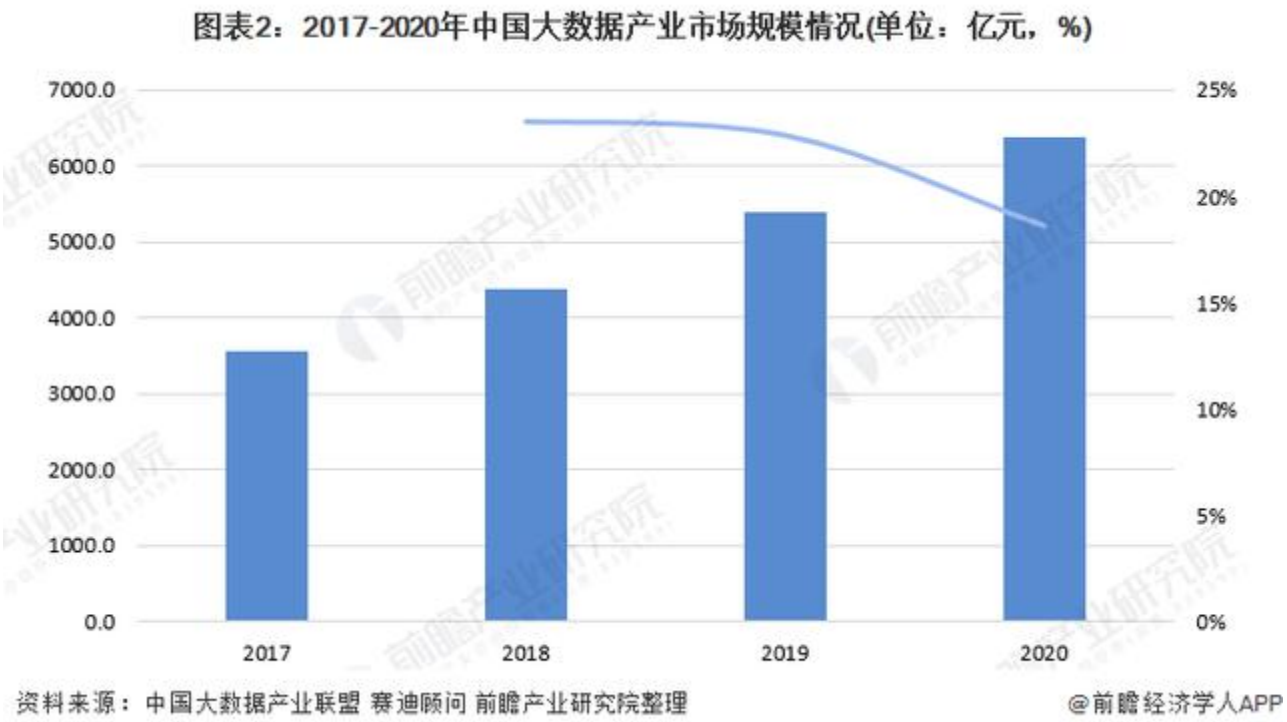
- 校园安防系统-数据预算

800台摄像头，支持1080分辨率的流媒体，一天的数据量？

假设摄像头为fps = 16;
1080分辨率为1920*1080bit
1920*1080*16 ？1秒钟的数据量？ 约为3.9MBps
压缩：H.264 250倍 H.265 300倍左右
3.9*60*60*24/300 约为1.1GB
实际情况：存在支持2K，4K的摄像头，他们的码流达到8MBps等

2TB~6TB/day

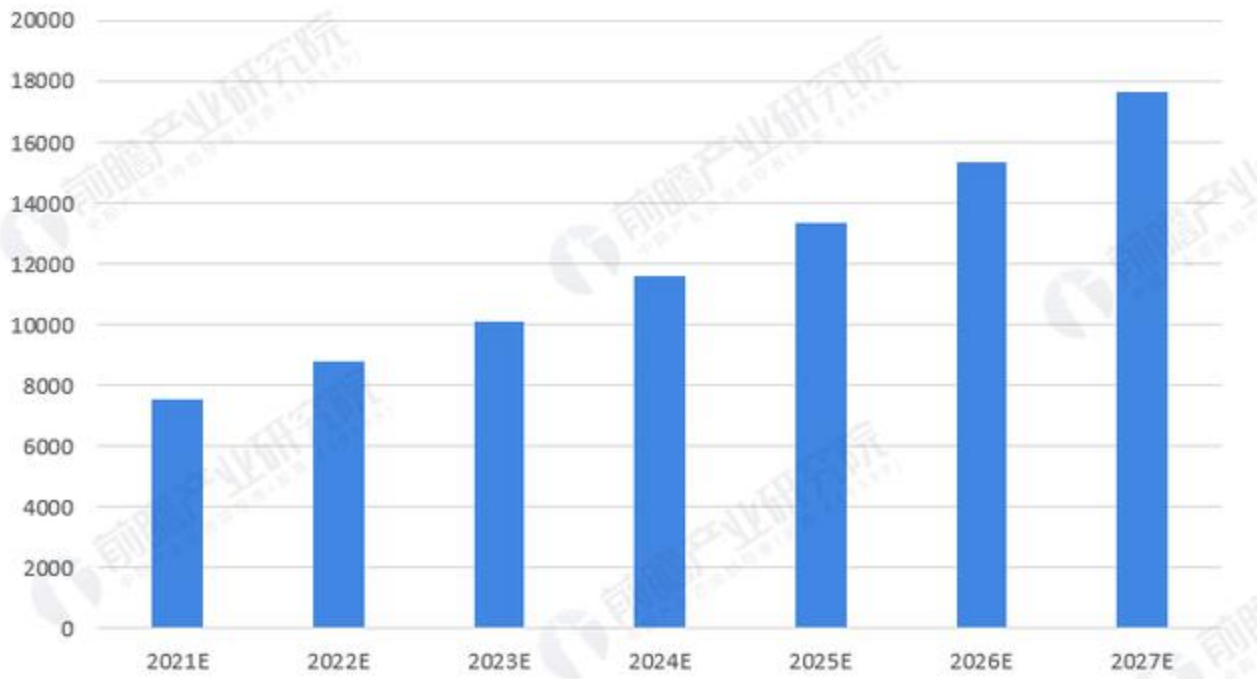
大数据时代-国内市场规模



2020年国内大数据产业市场规模超过6000亿人民币

大数据时代-大数据相关产业市场规模

图表6：2021-2027年中国大数据产业市场规模预测(单位：亿元)



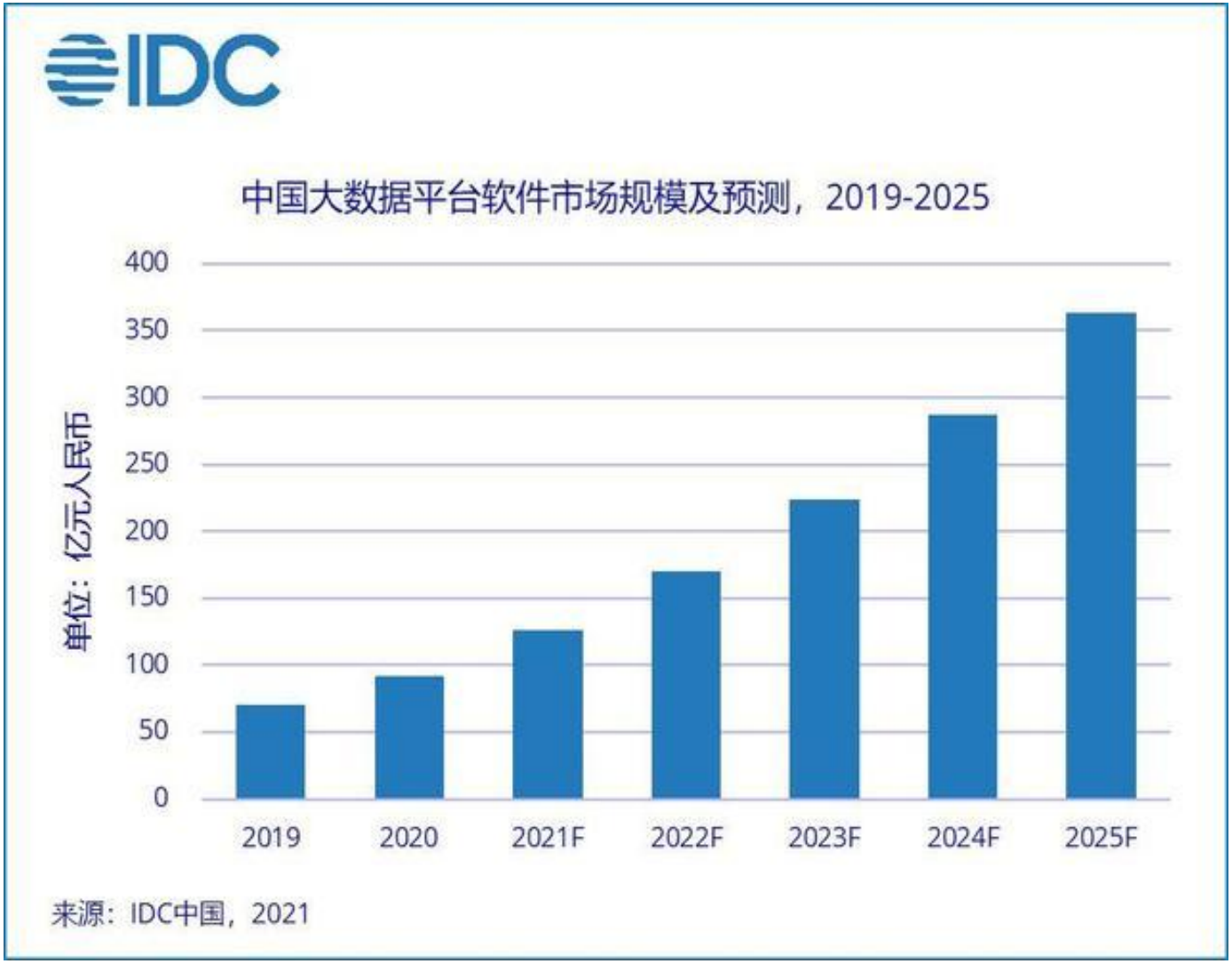
资料来源：前瞻产业研究院整理

@前瞻经济学人APP

预测2023年国内大数据产业市场规模超过一万亿人民币

根据光明网报道，2021年全年国内生产总值达1143670亿元

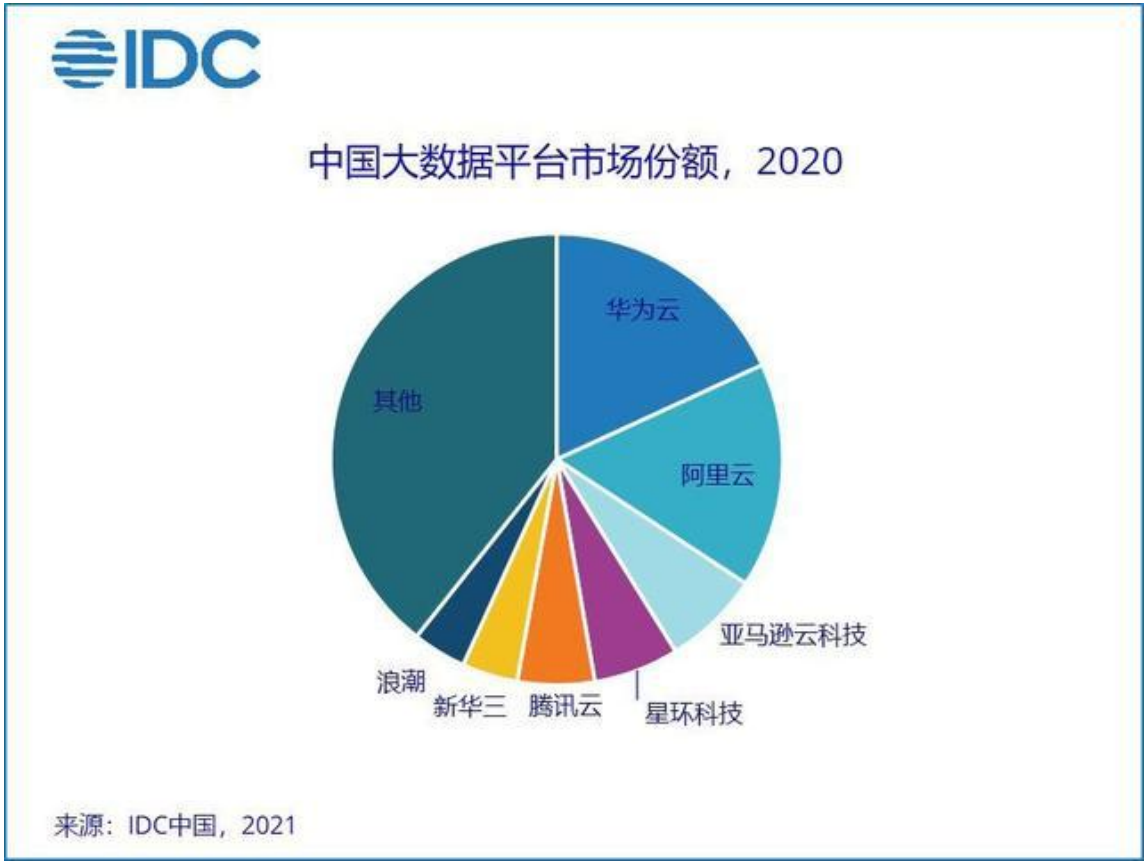
大数据时代-大数据平台软件市场规模



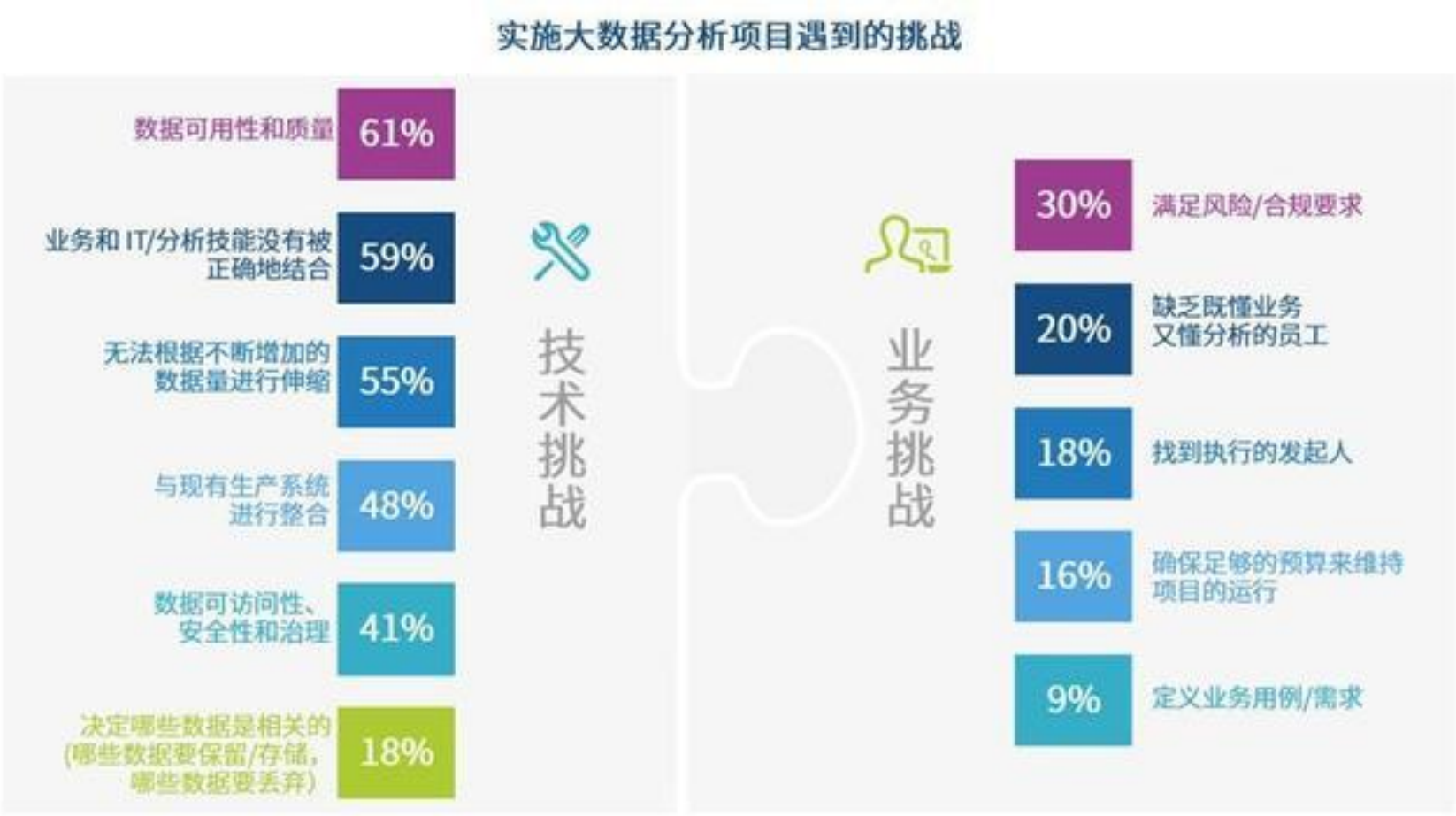
2021年国内大数据软件市场规模约为130亿人民币

全球大数据软件市场规模约为640亿美元

大数据时代-国内大数据平台市场份额



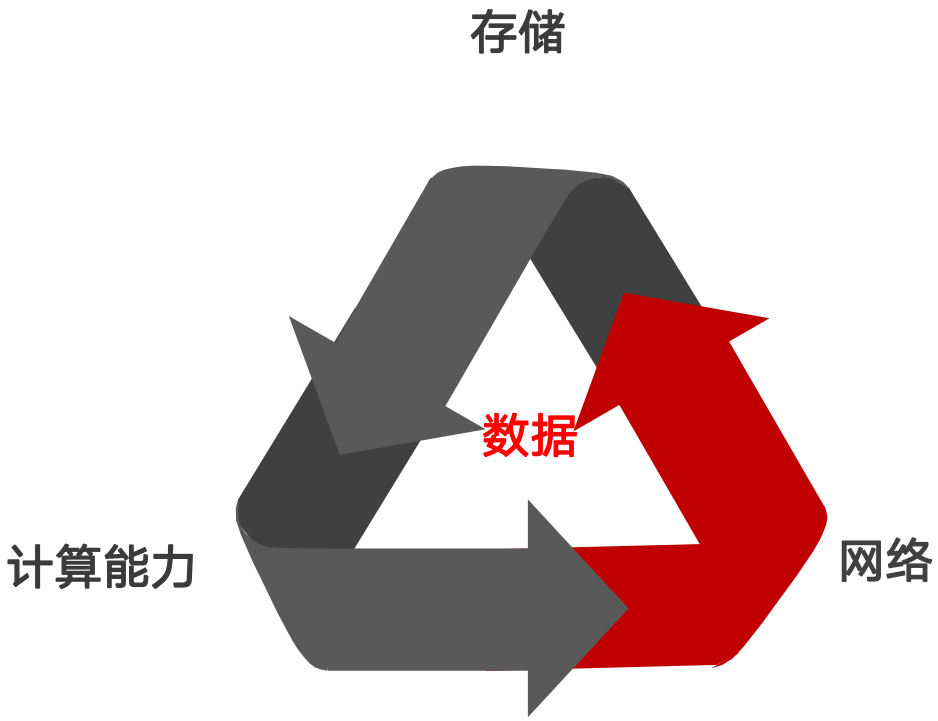
大数据项目实施挑战-技术&业务



n = 50
来源：IDC 亚太地区软件调查，2019

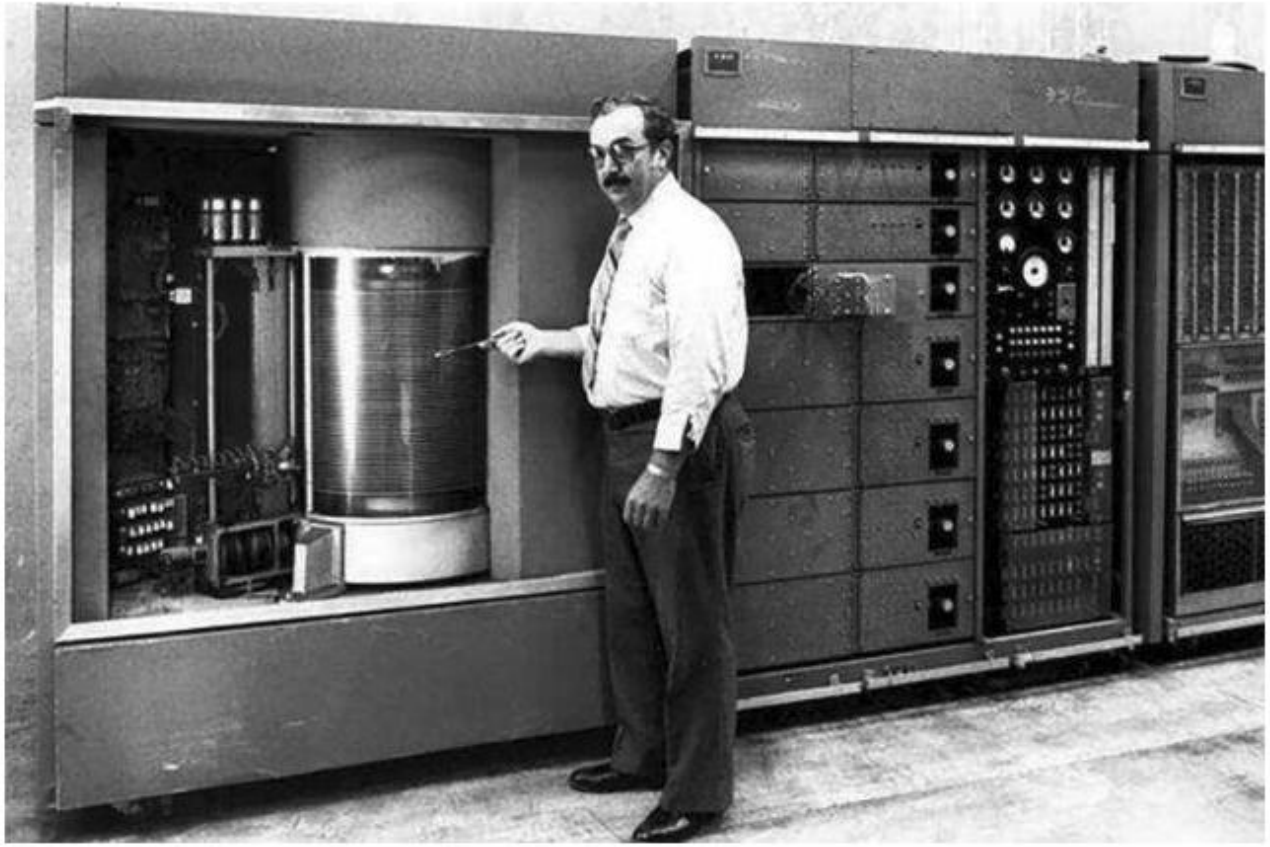
大数据时代的基石

有了爆炸增长的数据，我们还需要什么才能迎来大数据时代？



大数据时代的基石-存储

1956年，IBM发明了世界第一个HDD：350RAMAC。这个硬盘有50个24英寸盘片，只有5MB，但是却有两台冰箱一样大，重量超过一吨，被用于当时的工业领域。虽然350RAMAC还不能称之为严格意义上的硬盘，但却为计算机发展史掀起了新一页。



大数据时代的基石-存储

1991年，IBM推出来容量为1GB的3.5英寸硬盘——0663-E12，这款硬盘使用了MR磁头，使普通用户使用的硬盘容量首次达到了1GB，标志着硬盘进入GB时代。

2007年，日立推出其第一款突破TB级容量的硬盘，利用了垂直记录技术，在磁盘区域密度达到极限，磁盘容量受到超顺磁效应限制的情况下，实现了磁盘容量的进一步提升。

随着固态硬盘的兴起，机械硬盘似乎大势已去，而且受制于物理规则，机械硬盘的瓶颈难以突破，从2012年起，机械硬盘的发展几乎停滞了，各大厂商也纷纷将注意力转向固态硬盘，但近年来，随着存储需求增大，又有人打起了机械硬盘的主意。毕竟和固态硬盘相比，机械硬盘存在着价格和容量上的优势。

2006年3月，三星率先发布一款32GB容量的固态硬盘笔记本电脑，

2007年6月，东芝推出了其第一款120GB固态硬盘笔记本电脑。

2010年2月，镁光发布了全球首款SATA 6Gbps接口固态硬盘，突破了SATAII接口300MB/s的读写速度。

2013年SSD闪存技术取得长足的进步，在家用领域终于看到TB级SSD（此前最大的容量仅512GB）价格在3500-4000人民币之间。

大数据时代的基石-存储

京品电脑 希捷(Seagate)2TB 256MB 7200RPM 台式机机械硬盘 SATA接口 希捷酷鱼BarraCuda系列 (ST2000DM008)

【超级品类日！希捷酷鱼1/2/4/6TB到手价269/359/529/799！】可靠品质，优化读写，性能可靠【立即抢购】查看>

品牌闪购 距离结束 12 : 21 : 58

闪购价 ¥359.00 降价通知 累计评价 200万+

配送至 河南驻马店市泌阳县城区 可配货 支持 可配送海外 99元免基础运费

再次追平希捷 西数2T今降10元售840元

https://www.pcpop.com/article/553402.shtml

火狐官方站点 新手上路 常用网址 京东商城 生活 开发 microsoft 爬虫 论文 android goog webs Spring XX 开源镜像 cv

再次追平希捷 西数2T今降10元售840元

2010年07月12日 00:15 | 作者：宋阳 | 编辑：宋阳 | 文章出处：泡泡网原创 分享

泡泡网硬盘频道7月12日 希捷2TB昨天跌到840元，今天西数也追了上来，降到840元，两款产品的价格战大有愈演愈烈的趋势，不过受益者始终是消费者，相信大家很乐意看这种情况吧。

这款硬盘接口为SATA 3Gbps、转速7200RPM、缓存32MB、平均故障间隔时间100万小时、尺寸147×101.6×25.4毫米、重量690克、待机抗冲击力250G、工作抗冲击力65G等等，而且由于隶属于绿盘

大数据时代的基石-存储

1TB固态硬盘上市，价格是1TB机械硬盘的40倍

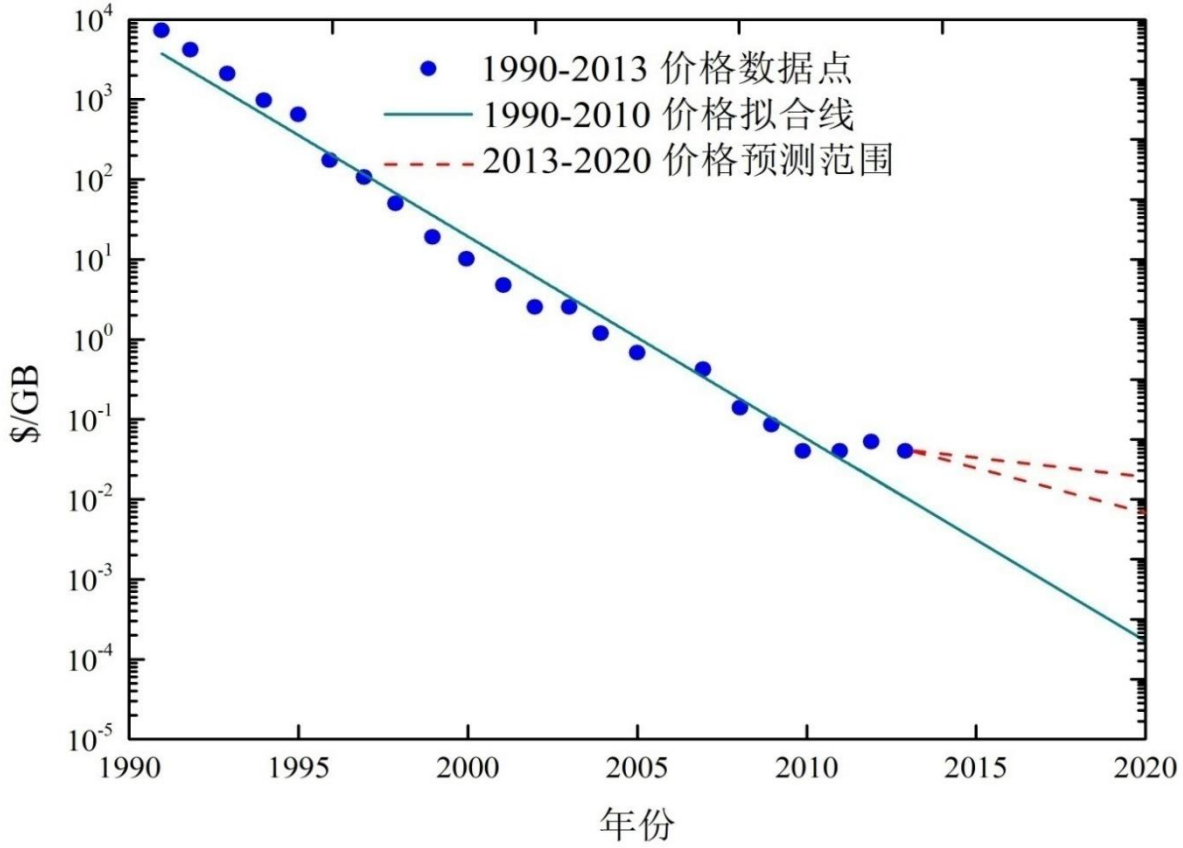
2010-04-14 mtjs 阅 127 转 1 分享 全屏 转藏

过去几年，固态硬盘（SSD）已逐渐从一个奢侈品变成了用户可以承担得起的消费品。但普通的机械式硬盘在容量上仍然有着绝对的优势，一般的SSD容量只有128GB，而普通硬盘超过了1TB。

> 现在，随着OCZ推出容量最高达1TB的Colossus LT系列固体硬盘的推出，这种情况发生了改变。Colossus LT系列SSD有120GB、250GB、500GB和1TB (1024GB) 四种型号。虽然普通硬盘的容量优势不再存在，但在性价比上SSD挑战普通硬盘仍有很长的路要走——1TB SSD售价高达4000美元，是1TB机械硬盘的40倍。

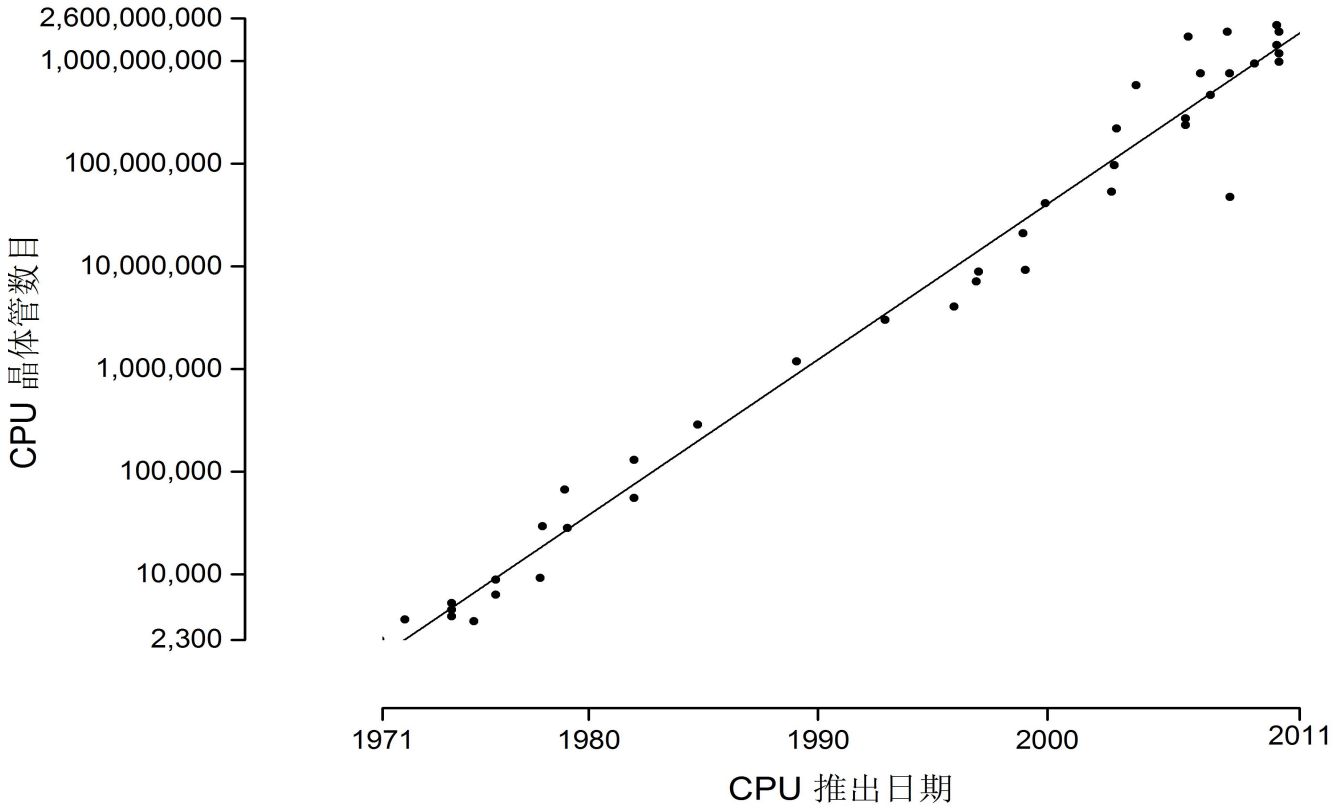
大数据时代的基石-存储

存储设备容量不断增加



大数据时代的基石-计算能力

CPU处理能力大幅提升



大数据时代的基石-计算能力

GPU加速数据处理

使用GPU进行计算过程加速已经成为学术界和产业界研究的主流，业务领域有成功的应用案例，如深度学习、密码分析、基因测序、入侵检测以及数据库系统优化等。GPU为应用开发提供了易用且性能强大的并行计算能力，极大地提高了各类计算密集型应用的性能。将GPU用于MapReduce模型优化，也已经取得了一系列的成果。

CUDA® 是 NVIDIA 专为图形处理单元 (GPU) 上的通用计算开发的并行计算平台和编程模型。借助 CUDA，开发者能够利用 GPU 的强大性能显著加速计算应用。

在经 GPU 加速的应用中，工作负载的串行部分在 CPU 上运行，且 CPU 已针对单线程性能进行优化，而应用的计算密集型部分则以并行方式在数千个 GPU 核心上运行。使用 CUDA 时，开发者使用主流语言（如 C、C++、Fortran、Python 和 MATLAB）进行编程，并通过扩展程序以几个基本关键字的形式来表示并行性。

NVIDIA 的 CUDA 工具包提供了开发 GPU 加速应用所需的一切。CUDA 工具包中包含多个 GPU 加速库、一个编译器、多种开发工具以及 CUDA 运行环境。

立即下载 ›

通过 CUDA 开发的数千个应用已部署到嵌入式系统、工作站、数据中心和云中的 GPU。

Adobe　Ansys　AUTODESK　DASSAULT SYSTEMES

MathWorks　Microsoft　NATIONAL INSTRUMENTS　Wolfram Mathematica

大数据时代的基石-网络

网络带宽不断增加

90年代中期至2000年：电话线+“猫” 龟速时代　　几十Kbps

2000-2006：ADSL(2+)宽带异军突起　　最高1Mbps上行8Mbps的下行速度

光纤入楼，光纤入户　　百兆光纤，千兆光纤

蜂窝网络，3G网络，4G网络，5G网络　　文字、声音、图片、视频……

大数据时代的基石-网络

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。

数据产生方式的变革促成大数据时代的来临

1 运营式系统阶段

- 数据库的出现使得数据管理的复杂度大大降低，数据往往伴随着一定的运营活动而产生并记录在数据库中，数据的产生方式是被动的

2 用户原创内容阶段

- 数据爆发产生于 Web 2.0 时代，而 Web 2.0 的最重要标志就是用户原创内容
- 智能手机等移动设备加速内容产生
- 数据产生方式是主动的

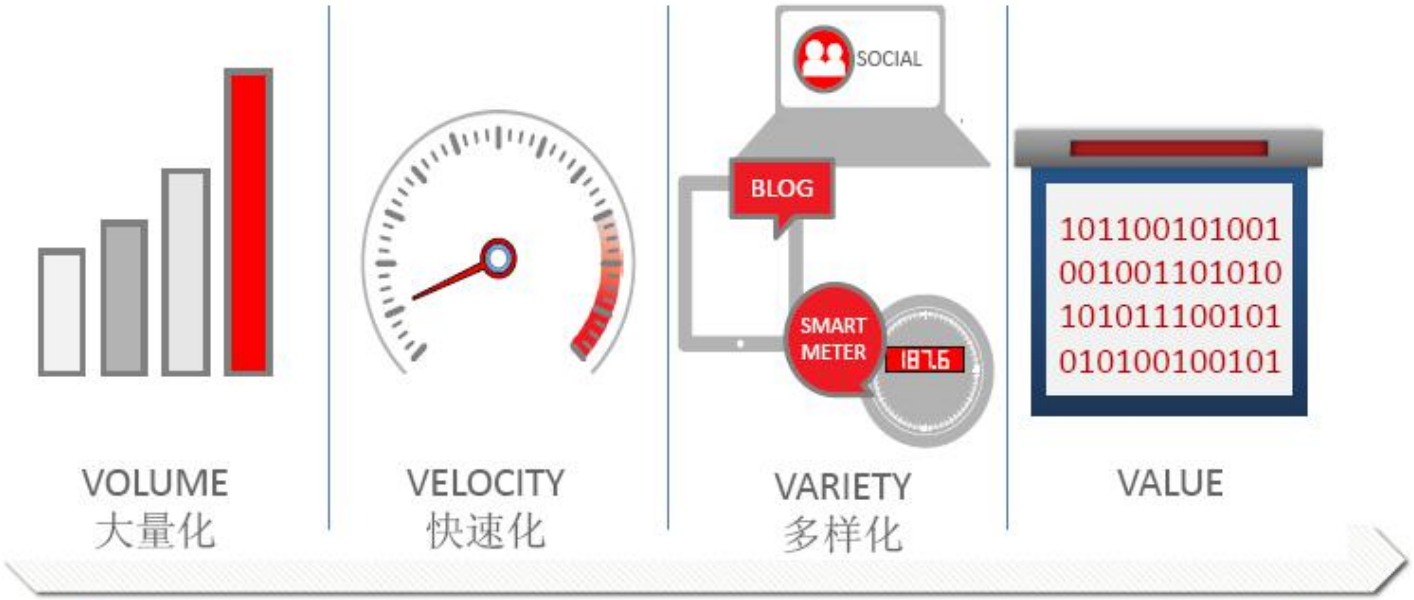
3 感知式系统阶段

- 感知式系统的广泛使用
- 人类社会数据量第三次大的飞跃最终导致了大数据的产生

大数据的发展历程

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2003年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

1.3大数据概念特点

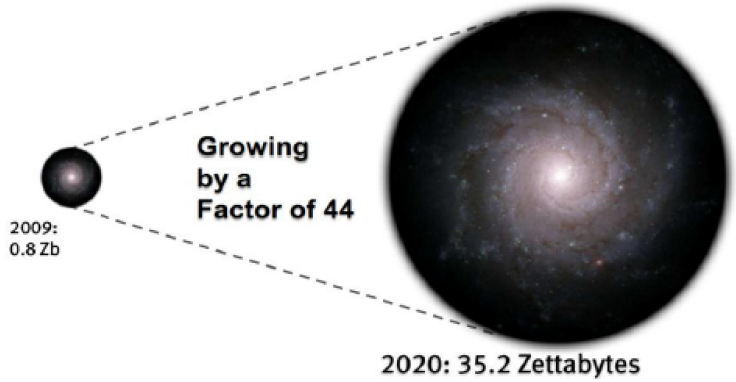


大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

4-V

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
- 到2020年，全球总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍
- 2030年，全球每年新增数据量将突破1YB量级，相当于4万亿台256GB高端手机的存储能力。

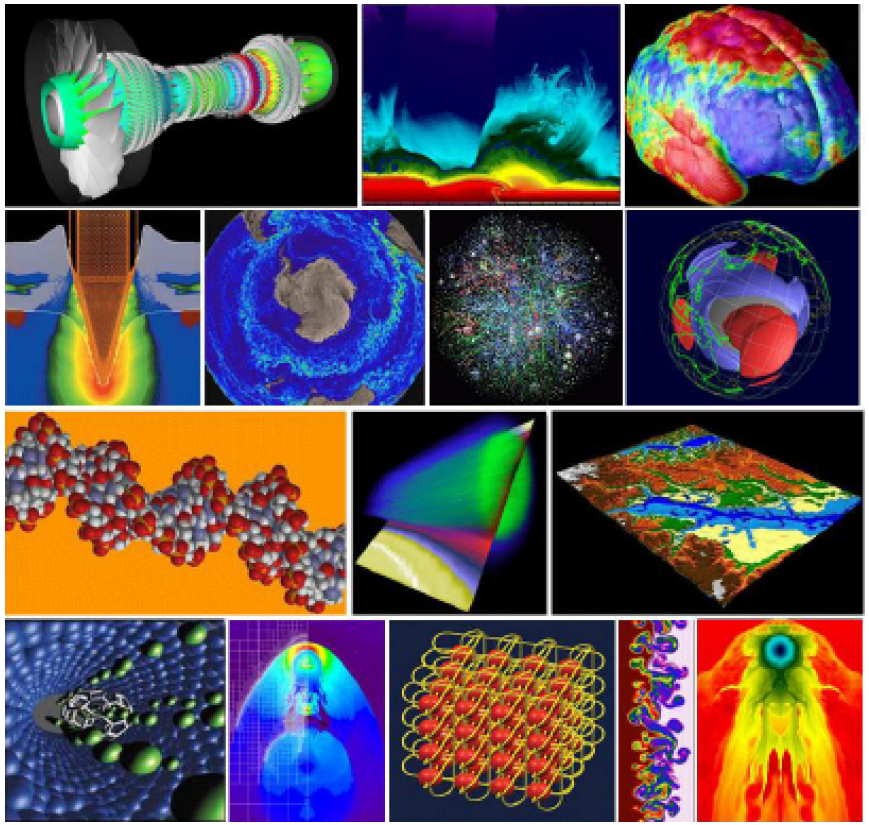


TERABYTE	10 的 12 次方	一块 1TB 硬盘		200,000 照片或 mp3 歌曲
PETABYTE	10 的 15 次方	两个数据中心机柜		16 个 Blackblaze pod 存储单元
EXABYTE	10 的 18 次方	2,000 个机柜		占据一个街区的 4 层数据中心
ZETTABYTE	10 的 21 次方	1000 个数据中心		纽约曼哈顿的 1/5 区域
YOTTABYTE	10 的 24 次方	一百万个数据中心		特拉华州和罗德岛州

微软将 27.6PB 的数据，2018 年被沉入了苏格兰奥克尼群岛的一片海域

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



- 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
- 流媒体数据
  - 抖音、快手等短视频，直播数据

数据类型繁多

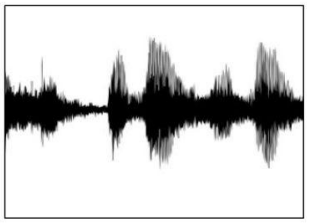
相对于结构化数据（即行数据，存储在数据库里，可以用二维表结构来逻辑表达实现的数据）而言，不方便用数据库二维逻辑表来表现的数据即称为非结构化数据，包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等。

Structured Data

Size	#bedrooms	···	Price (1000$s)
2104	3		400
1600	3		330
2400	3		369
⋮	⋮		⋮
3000	4		540

User Age	Ad Id	···	Click
41	93242		1
80	93287		0
18	87312		1
⋮	⋮		⋮
27	71244		1

Unstructured Data



Audio

Image

Four scores and seven years ago…

Text

数据类型繁多

下面的众多数据中，哪些是结构化数据，哪些是非结构化数据？

信用卡号码、日期、财务金额、天气数据、地形、军事活动、石油和天然气勘探、空间勘探、地震图像、大气数据、监控照片和视频、交通信息、天气、海洋传感器、电话号码、地址、产品名称、文本文件、电子邮件、微博数据、网站数据、通讯数据、媒体、ppt课件数据

数据类型繁多

典型的结构化数据：

信用卡号码、日期、财务金额、电话号码、地址、产品名称等；

典型的非结构化数据：

文本文件：文字处理、电子表格、演示文稿、电子邮件、日志。
电子邮件：电子邮件由于其元数据而具有一些内部结构，我们有时将其称为半结构化。但是，消息字段是非结构化的，传统的分析工具无法解析它。
社交媒体：新浪微博、微信、QQ、Facebook的数据。
网站： YouTube，Instagram，照片共享网站。
移动数据：短信、位置等。
通讯：聊天、即时消息、电话录音、协作软件等。
媒体：MP3、数码照片、音频文件、视频文件。
卫星图像：天气数据、地形、军事活动。
科学数据：石油和天然气勘探、空间勘探、地震图像、大气数据。
数字监控：监控照片和视频。
传感器数据：交通、天气、海洋传感器。

处理速度快

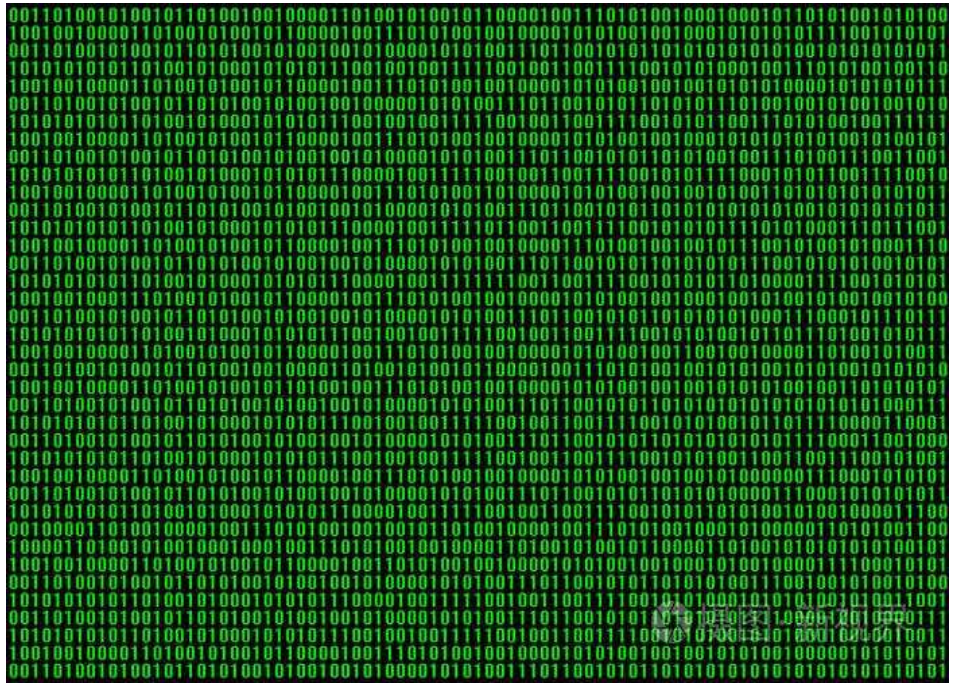
- 从数据的生成到消耗，时间窗口非常小，可用于生成决策的时间非常少
- 1秒定律：互联网进入DT时代对数据处理速度的要求，也不是就要求到1秒钟之内，而一般为秒级的响应。这一点也是和传统的数据挖掘技术有着本质的不同

比如淘宝的推荐功能

价值密度低

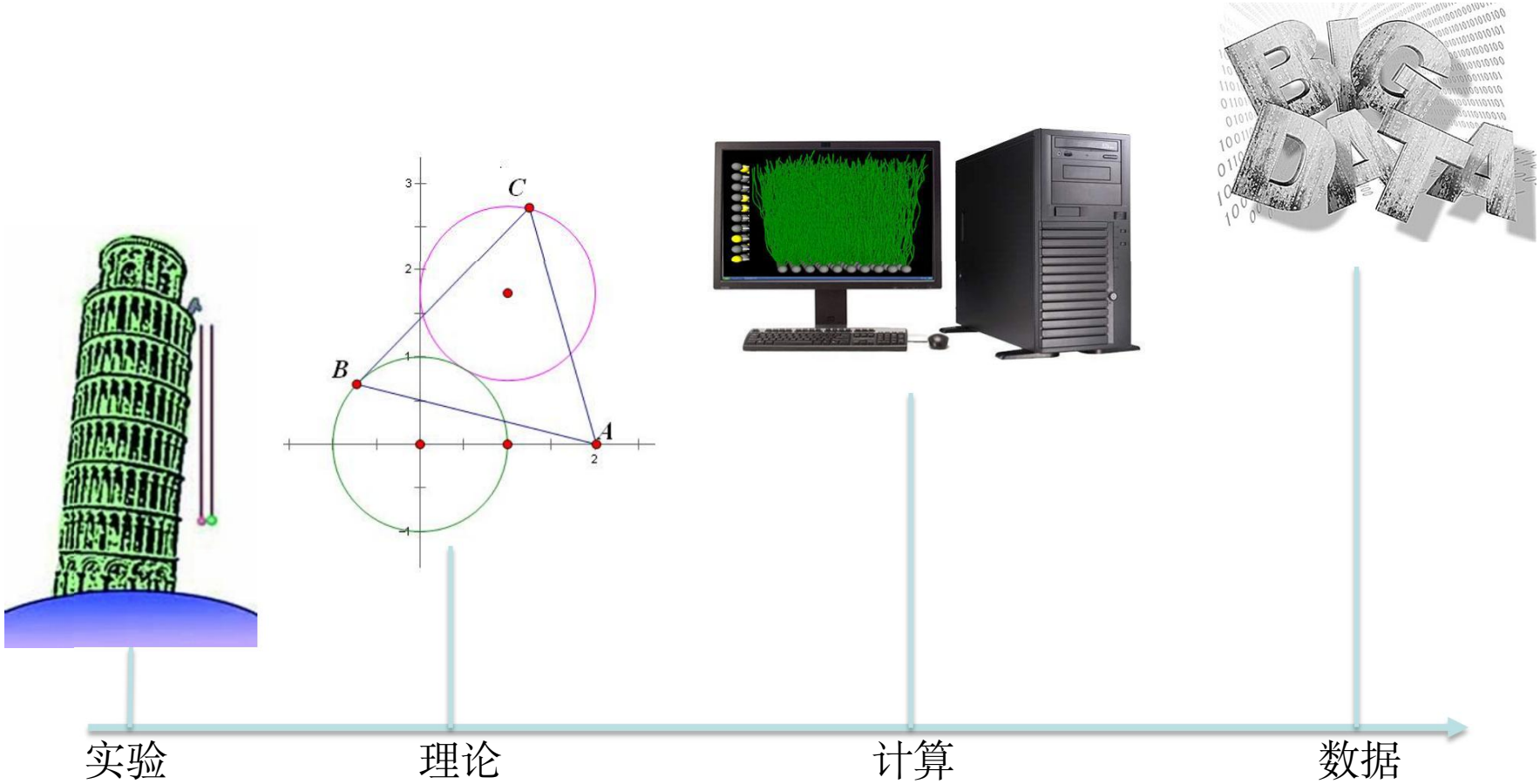
价值密度低，商业价值高

以学校监控视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的价值。



1.4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



1.4大数据的影响

举个例子：

NLP（自然语言处理），也就是我们通常所说的语音识别相关领域。

1947年，美国科学家韦弗（W. Weaver）博士和英国工程师布斯（A. D. Booth）提出了利用计算机进行语言自动翻译的设想，机器翻译（Machine Translation）从此步入历史舞台；

1957 年，麻省理工学院的语言学教授诺姆·乔姆斯基（Noam Chomsky）在他出版的《句法结构》一书中，革新了语言的概念，提出“要使计算机理解语言，就必须更改句子的结构。”

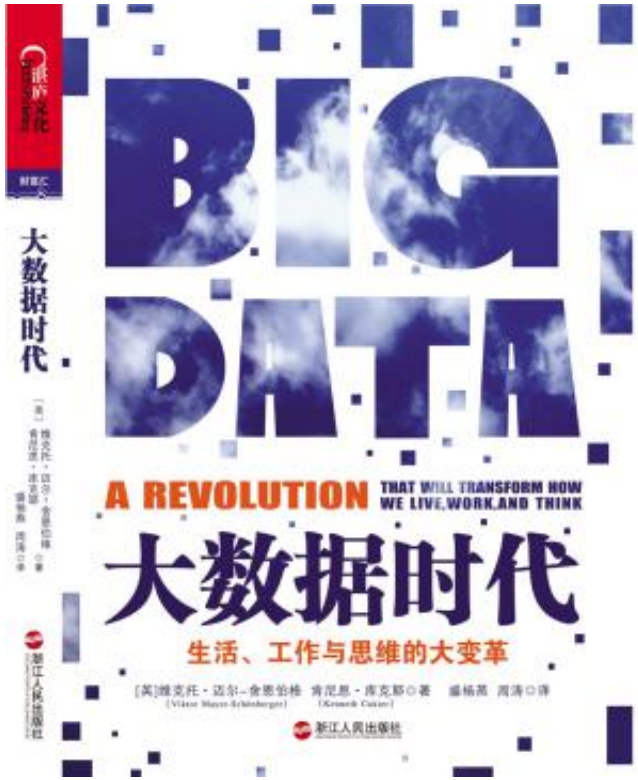
但是词法，句法太过发杂，可以重组的可能太多，做出的产品无法达到要求且成本巨大。

早期的自然语言处理研究中，很流行语言学和统计学的混合，大多数 NLP 系统都使用复杂的“手写”逻辑规则。而现在，这一理念被**纯粹的统计学**所取代。20 世纪 80 年代，得益于计算能力的稳定增长以及机器学习的发展，研究人员开始对 AI 和 NLP 进行根本性的重新定位，用简单的近似法取代了深入的分析法，评估过程也变得更加量化。

由结构规律算法到存粹统计学方法，以数据为解决方案而不再以词法、句法结构特点为唯一解决方案。

1.4大数据的影响

- 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**

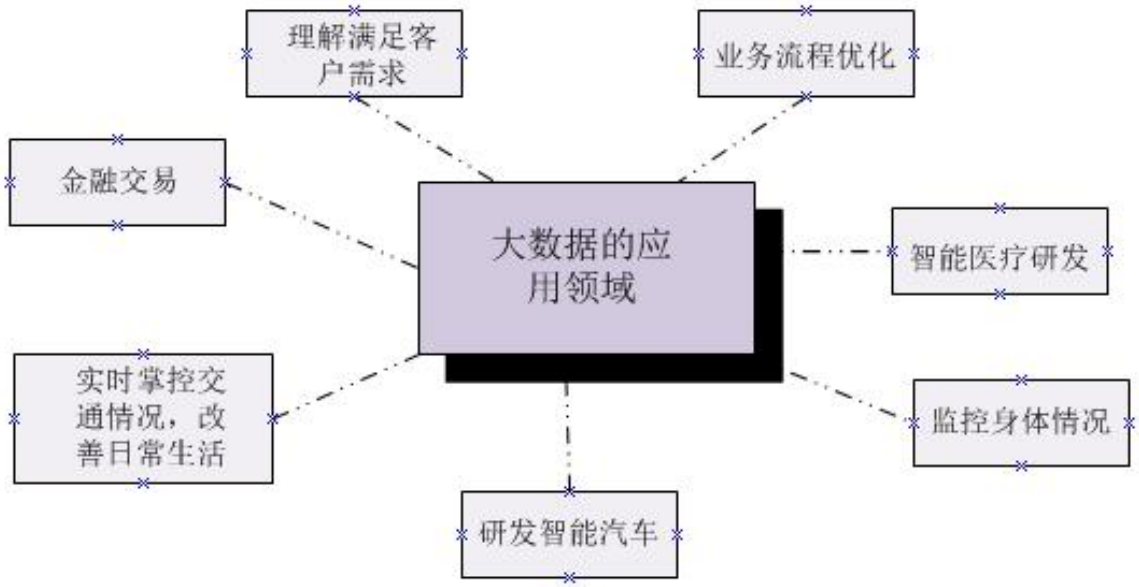


1.4大数据的影响

- 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

1.5大数据应用

- 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

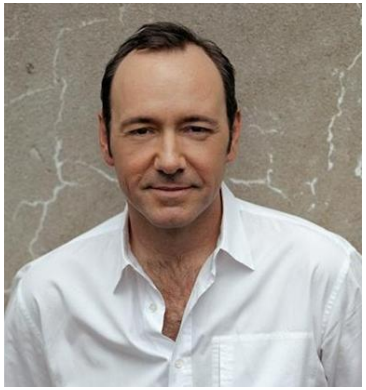


典型的大数据应用实例-金融

金融行业

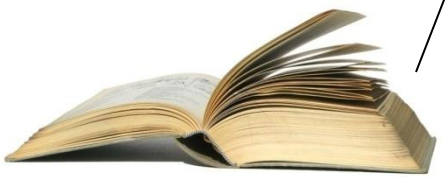
- **银行**：信用评分，放贷，投资风险预测等。
- **保险**：比如基于大量天气数据的分析可以为相关地方的农业保险定价。
- **证券**：通过爬取互联网上的信息进行分析来知道选股，比如百度百发、淘金100等都属于大数据基金。

典型的大数据应用实例-娱乐

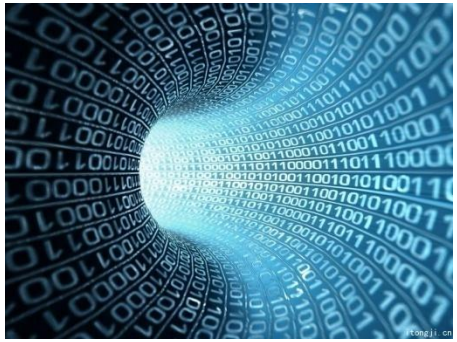


Kevin Spacey

David Fincher



英国同名小说《纸牌屋》



大数据分析

《纸牌屋》作为当时第一个在流媒体上播放的主流电视剧，一经播放万人空巷，横扫了当年的影视市场。

风靡全球的美剧《纸牌屋》

典型的大数据应用实例-社科

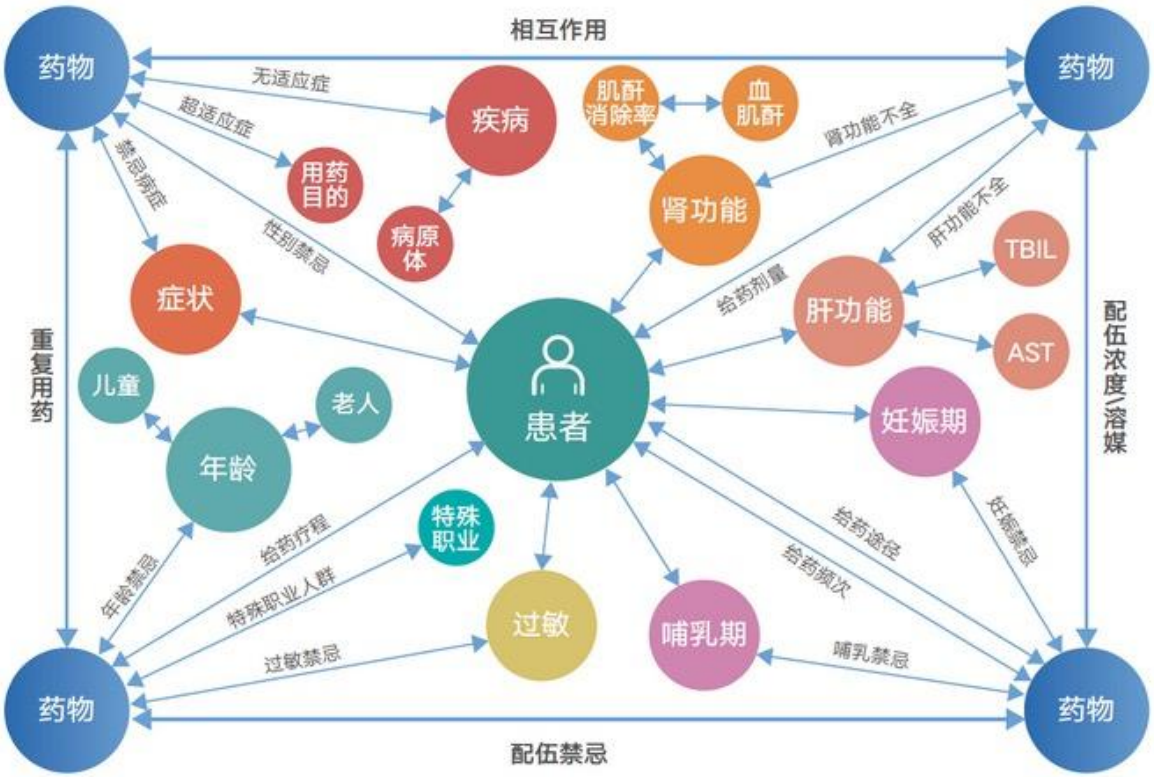
从谷歌流感趋势看大数据的应用价值

“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

典型的大数据应用实例-医疗

多场景应用

典型的大数据应用实例-医疗

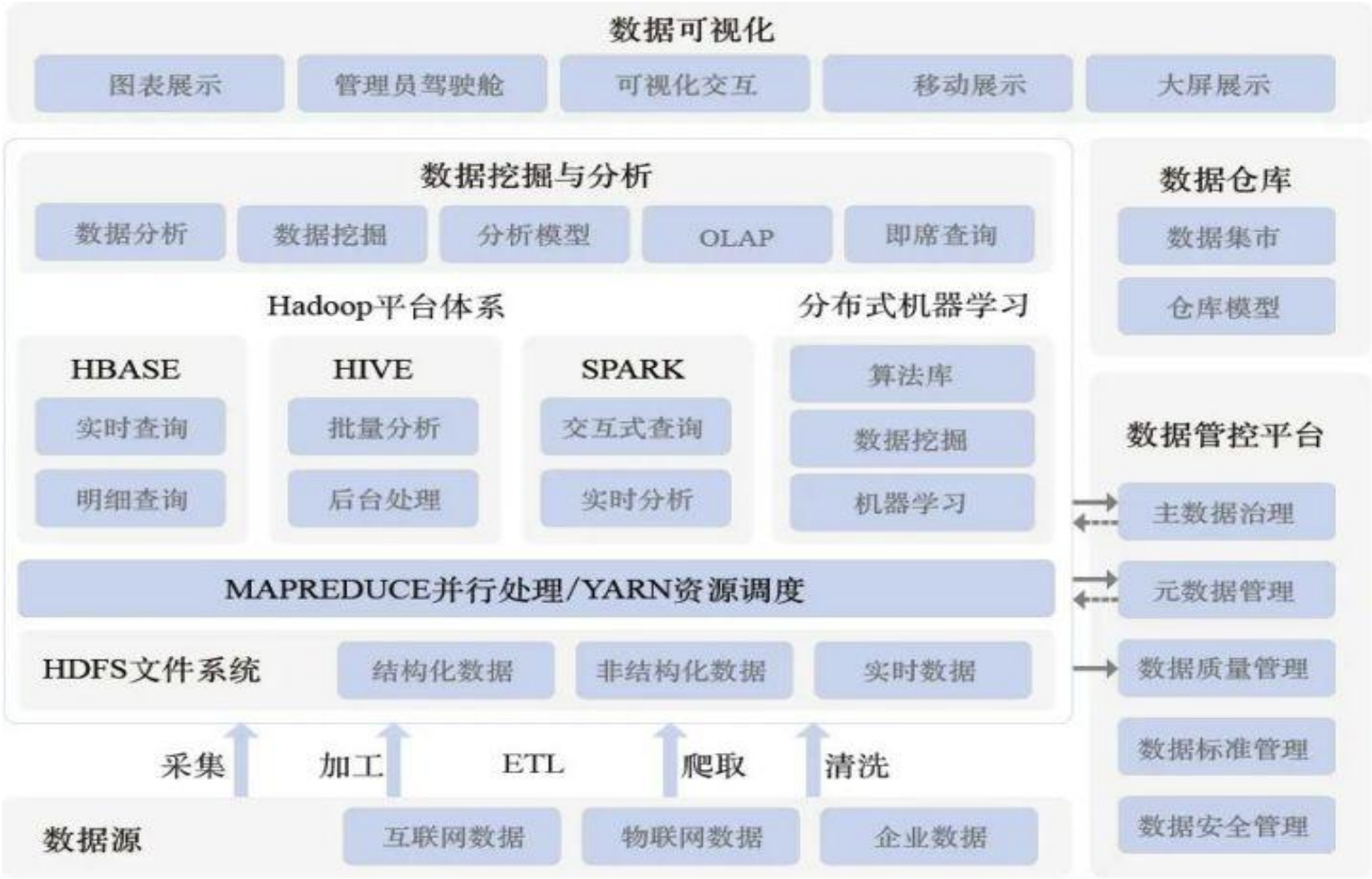


支持全部质控维度的知识查看，支持自主调整知识内容，提示内容及提示等级，帮助药师更高效的主动、自主审查药方的合规性。

1.6大数据关键技术

技术层面	功能
数据采集	利用ETL(Extract-Transform-Load，抽取-转化-加载)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

大数据应用整体架构


数据可视化
图表展示
管理员驾驶舱
可视化交互
移动展示
大屏展示
数据挖掘与分析
数据分析
数据挖掘
分析模型
OLAP
即席查询
Hadoop平台体系
分布式机器学习
HBASE
实时查询
明细查询
HIVE
批量分析
后台处理
SPARK
交互式查询
实时分析
算法库
数据挖掘
机器学习
MAPREDUCE并行处理/YARN资源调度
HDFS文件系统
结构化数据
非结构化数据
实时数据
采集
加工
ETL
爬取
清洗
数据源
互联网数据
物联网数据
企业数据
数据仓库
数据集市
仓库模型
数据管控平台
主数据治理
元数据管理
数据质量管理
数据标准管理
数据安全管理

大数据管理数据处理流程

- 从“做饭”的程序来讲大数据
- 【主菜】
- 正所谓“巧妇难为无米之炊”，做饭首先得有食材，大数据也一样，没有数据说什么都是空谈，所以数据就是数据人的食材（只要有数据，我不用吃饭）。

- 做饭通常都要包括“买菜~洗菜~配菜~炒菜”这几个必须环节，无论你是开饭店还是家里一日三餐，做饭的规模大小会有不同，但流程却是一样的。而这几个环节其实正好对应了数据人的日常工作内容：买菜（数据采集）~洗菜（数据清洗）~配菜（数据建模）~炒菜（数据加工）。

大数据管理数据处理流程

菜市场就是我们通常所说的数据源

超市——信息化数据

数据种类有限，但通常是关键数据，数据格式规范

农贸市场——物联网数据

数据种类丰富，选择多，但品质参差不齐，数据格式多变

早市——外部数据

数据种类五花八门，价格便宜，但含金量低，后期处理成本高

大数据管理数据处理流程

数据采集就是一个买菜过程

选菜&砍价——数据调研&数据接口

选择需要的数据，制定接口规范

提菜回家——数据传输

打通数据传输链路，保证数据传输效率

数据坛

大数据管理数据处理流程

数据就跟菜一样，也是需要清洗的

数据质量检查

根据规则，检查不规范数据

数据校对

比对数据采集、计算结果

数据标准化

按统一标准对数据进行归并

大数据管理数据处理流程

数据建模跟配菜一样，都是个艺术活

按食材分类——范式建模

Bill Inmon喜欢吃火锅

按菜品分类——维度建模

Ralph Kimball喜欢吃炒菜

数据烩

大数据管理数据处理流程

数据加工就好比炒菜，是为最后的数据消费者服务的

满汉全席——仪表盘

可视化效果好，查询速度快
重要数据一目了然

自助餐——自助分析

数据丰富，基本覆盖各种需求
需要具备一定的动手能力

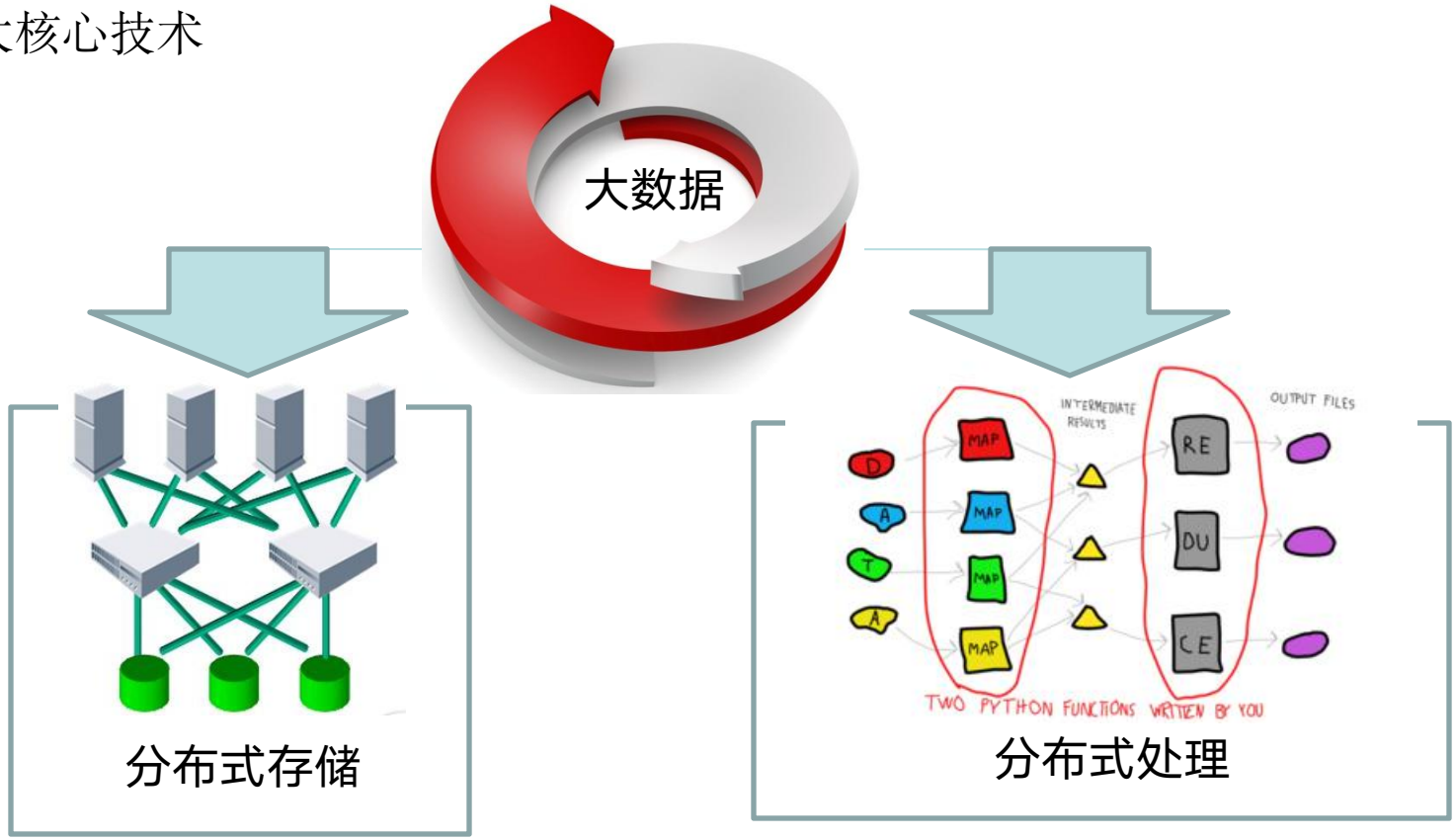
快餐——报表

数据内容少，但查询速度快
可大量供应

大数据管理数据处理流程

Hadoop平台核心技术

两大核心技术



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

MapReduce

数据获取手段-网络爬虫

爬虫框架

Scrapy框架

BeautifulSoup

Requests 库

pyspider

web Magic

爬虫框架就是一些爬虫项目的半成品。比如我们可以将一些常见爬虫功能的实现代码写好，然后留下一些接口，在做不同的爬虫项目时，我们只需要根据实际情况，手写少量需要变动的代码部分，并按照需要调用这些接口，即可以实现一个爬虫项目。

举个例子：

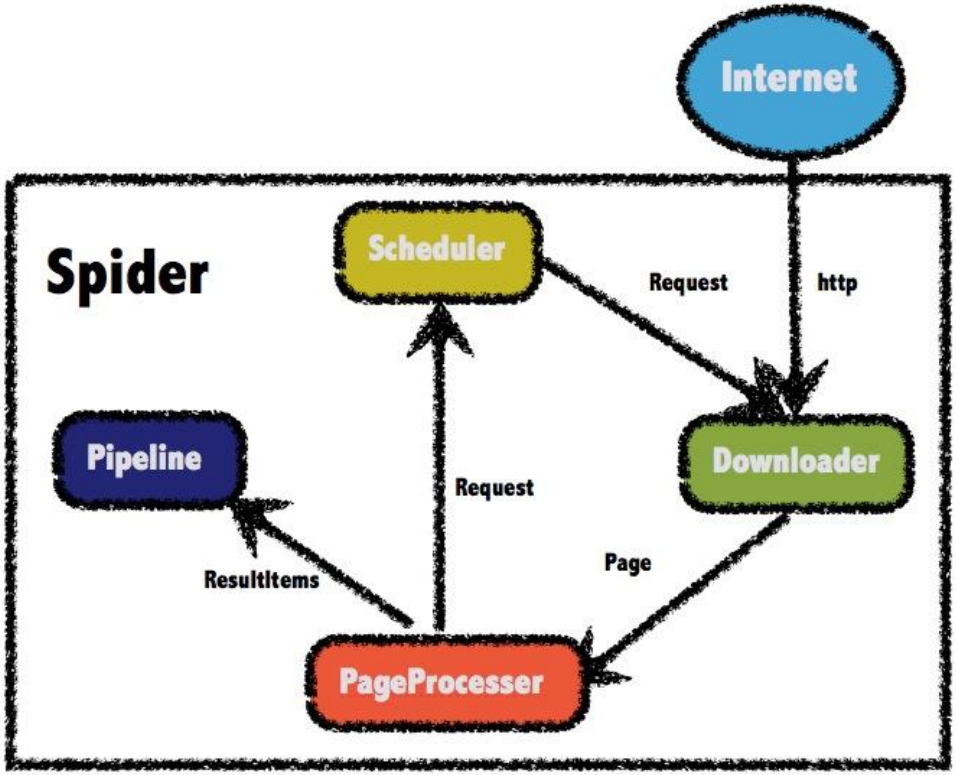
https://zmd.fang.anjuke.com/loupan/?pi=baidu-cpcaf-zmd-ty1&kwid=285985500143

数据获取手段-网络爬虫

```
▼<div class="list-contents theme-ajk-listcont">
  <input id="emailSubscribe-filter" value="p=1" type="hidden">
  ▼<div class="list-results">  event
    ▼<div class="key-sort">
      ▶<div class="page-tab">…</div>
      ▶<div class="sort-condi">…</div>
      </div>
      <!--优惠优盘入口-->
      <!--优惠楼盘置顶数据-->
      <!--团购2.0城市且有优惠楼盘推荐数据不展示聚客宝-->
      <div class="resource-ifx resource-ifx-filter" data-resource_id="1310310" data-region_id="" data-subregion_id="" data-city_id="172" data-soj="zybh-pc-31" data-limit="2"></div>
       event
    ▼<div class="key-list imglazyload">
      <!--2017/07/27-->
      ▼<div class="item-mod" data-link="https://zmd.fang.anjuke.com/loupan/456775.html" data-soj="AF_RANK_1" rel="nofollow">
        ▼<a class="pic" href="https://zmd.fang.anjuke.com/loupan/456775.html?from=AF_RANK_1" soj="AF_RANK_1" target="_blank">
          <i class="iconfont item-mod-video34 icon-mod-video35"></i>
          ▼<div class="icon-qj-weliao">
            <i class="iconfont title-icon icon-qj icon-quanjing"></i>
          </div>
          <img src="https://pic4.ajkimg.com/display/xinfang/b9f0ceb1ad6933c654a69b15f8766fd2/180x135m.jpg" alt="驻马店驿城区驿城区鹏宇城楼盘新房真实图片" width="180" height="135">
        </a>
        ▼<div class="infos">
          ▶<a class="lp-name" href="https://zmd.fang.anjuke.com/loupan/456775.html?from=AF_RANK_1" soj="AF_RANK_1" target="_blank">…</a>
          ▼<a class="address" href="https://zmd.fang.anjuke.com/loupan/456775.html?from=AF_RANK_1" soj="AF_RANK_1" target="_blank">
            <span class="list-map" target="_blank">[ 驿城区 驿城区 ] 置地大道与兴业大道交会处</span>
          </a>
          ▶<a class="huxing" href="https://zmd.fang.anjuke.com/loupan/456775.html?from=AF_RANK_1" soj="AF_RANK_1" target="_blank">…</a>
          ▶<a class="tags-wrap" href="https://zmd.fang.anjuke.com/loupan/456775.html?from=AF_RANK_1" soj="AF_RANK_1" target="_blank">…</a>
          <!--惠、抢、看-->
          ▶<div class="active-module">…</div>
        </div>
        ▼<a class="favor-pos" href="https://zmd.fang.anjuke.com/loupan/456775.html?from=AF_RANK_1" soj="AF_RANK_1" target="_blank">
          ▼<p class="price">
            均价
            <span>5300</span>
            元/㎡
          </p>
        </a>
```

网络爬虫基本流程

BeautifulSoup框架流程



WebMagic框架流程

数据分析&数据挖掘

（1）定义：简单来说，数据分析就是对数据进行分析。专业的说法，数据分析是指根据分析目的，用适当的统计分析方法及工具，对收集来的数据进行处理与分析，提取有价值的信息，发挥数据的作用。

（2）作用：它主要实现三大作用：现状分析、原因分析、预测分析。数据分析的目标明确，先做假设，然后通过数据分析来验证假设是否正确，从而得到相应的结论。

（3）方法：主要采用对比分析、分组分析、交叉分析、回归分析等常用分析方法；

（4）结果：数据分析一般都是得到一个指标统计量结果，如总和、平均值等，这些指标数据都需要与业务结合进行解读，才能发挥出数据的价值与作用。

数据分析&数据挖掘

数据挖掘是通过仔细**分析大量数据来揭示有意义的新的关系、趋势和模式**的过程.

数据挖掘的任务就是发现隐藏在数据中的模式.其可以发现的模式一般分为两大类:描述型(descriptive)模式和预测型(predictive)模式.描述型模式是对当前数据中存在的事实做规范描述,刻画当前数据的一般特性;预测型模式则是以时间为关键参数,对于时间序列型数据,根据其历史和当前的值去预测其未来的值.

分类模式 (Classification)

聚类模式(Clustering)

回归模式(Regression)

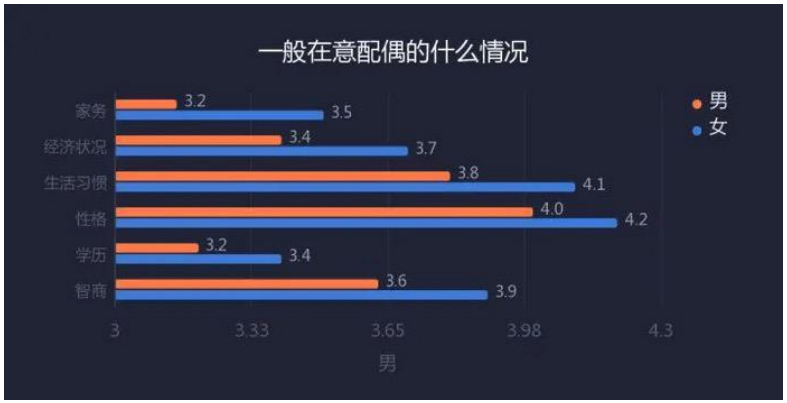
关联模式(Association)

序列模式[Sequential)

偏差模式(Deviation)

数据可视化

数据分析需要可视化
可视化可以带来更好的分析



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	Flink、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

1.8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

国内大数据产业全景图

大数据产业全景图
硬件设施
采集设备
传输设备
计算存储
设计集成
基础设施
网络服务
云平台服务
系统开发
数据来源
数据采集
政府大数据
工业大数据
技术开发
融合应用
工业
农业
政府
金融
医疗
电信
产业支撑
贵阳大数据交易所
武汉长江大数据交易中心
中关村大数据产业联盟
全国高校大数据教育联盟
资料来源：前瞻产业研究院整理

1.9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

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

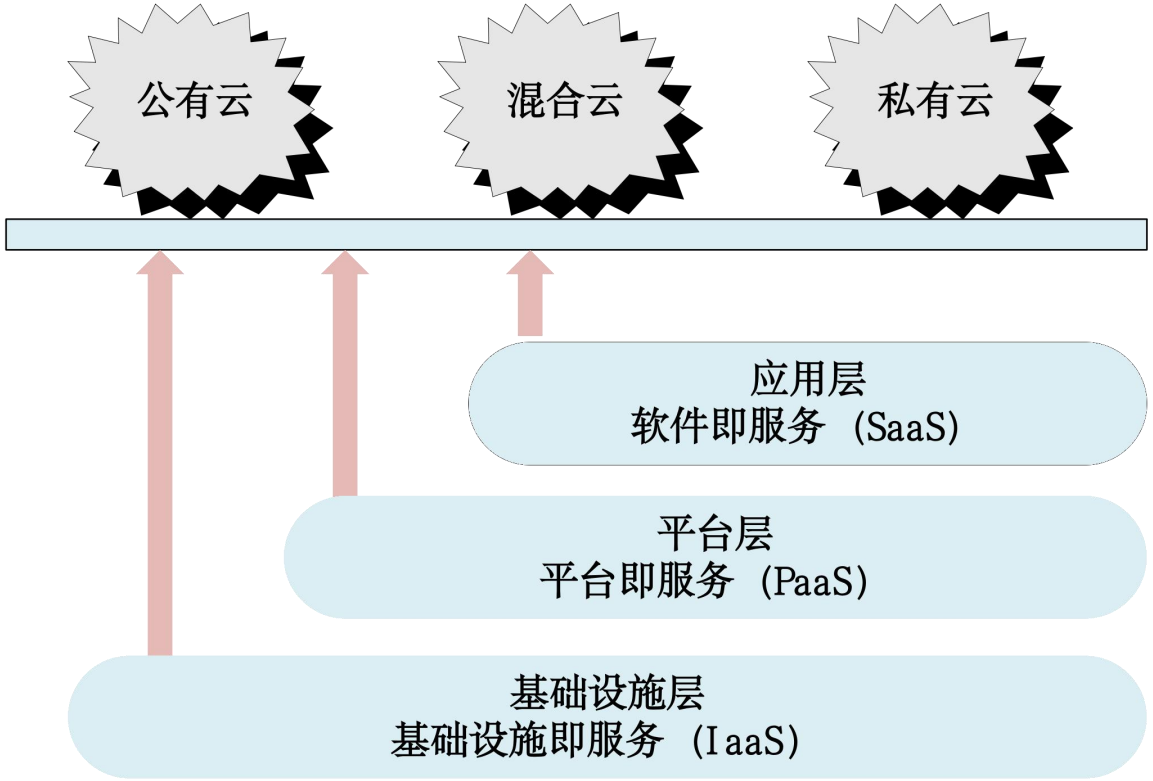
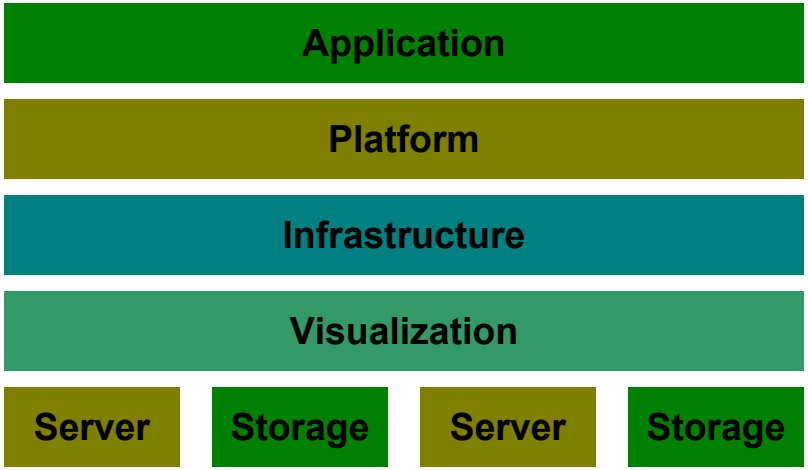


图1-7 云计算的服务模式和类型

云计算

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

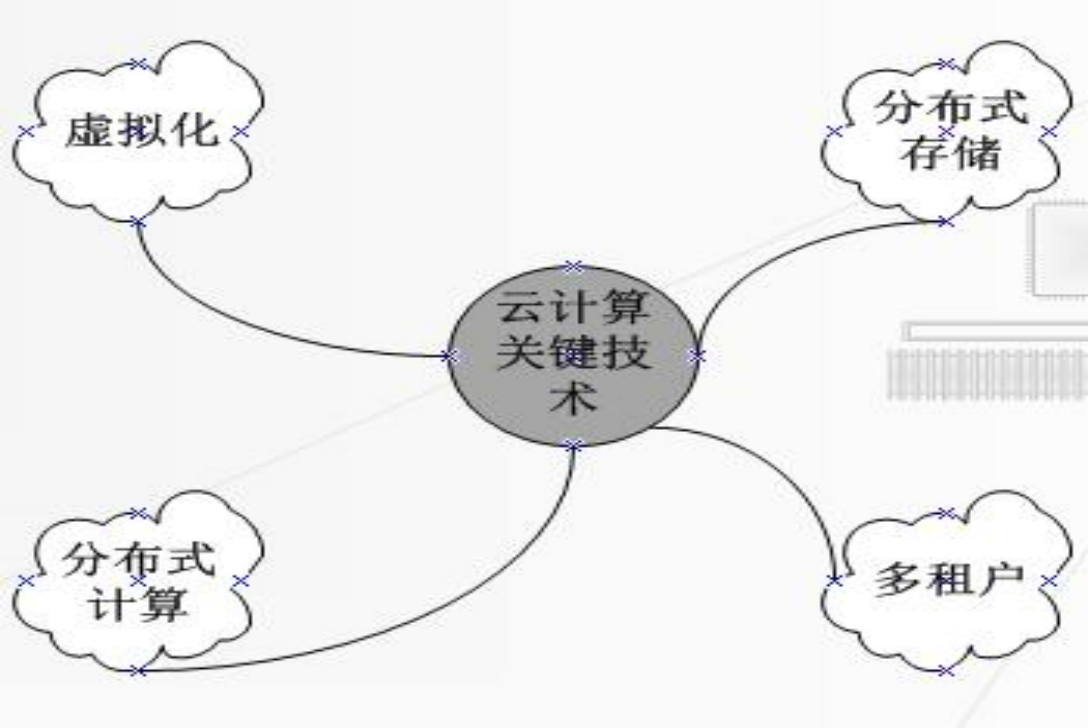


SaaS	Software as a Service	Google Apps, Microsoft “Software+Services”
PaaS	Platform as a Service	IBM IT factory, Google App Engine，Force.com
IaaS	Infrastructure as a Service	Amazon EC2, IBM Blue Cloud, Sun Grid

云计算

2. 云计算关键技术

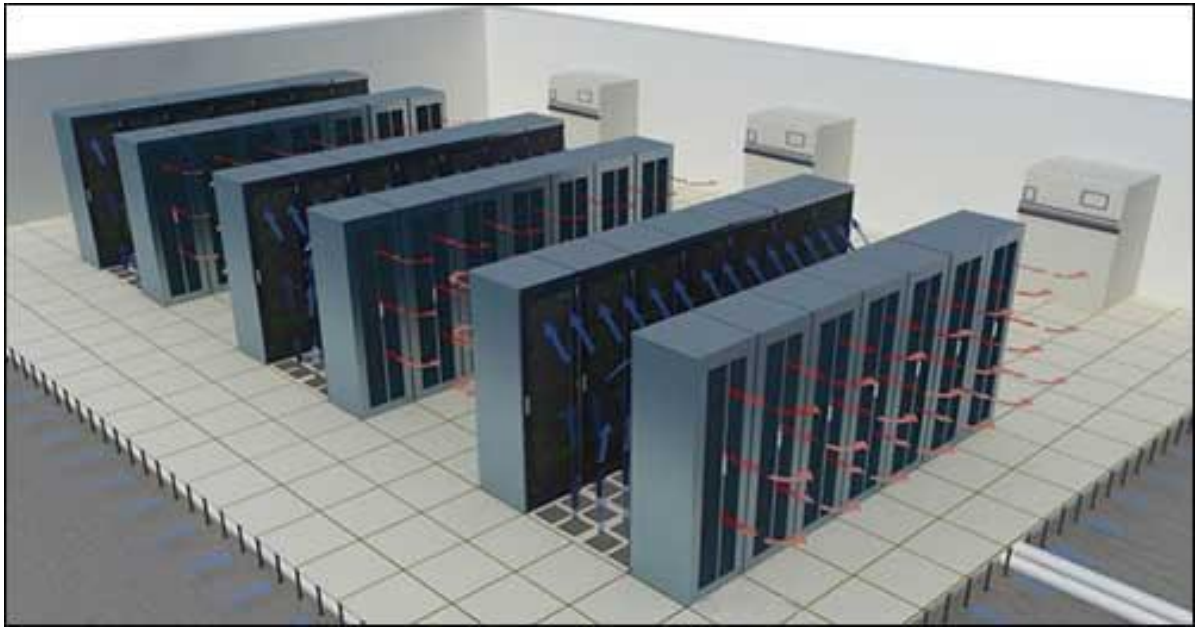
- 云计算关键技术包括：虚拟化、分布式存储、分布式计算、多租户等



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



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

云计算

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

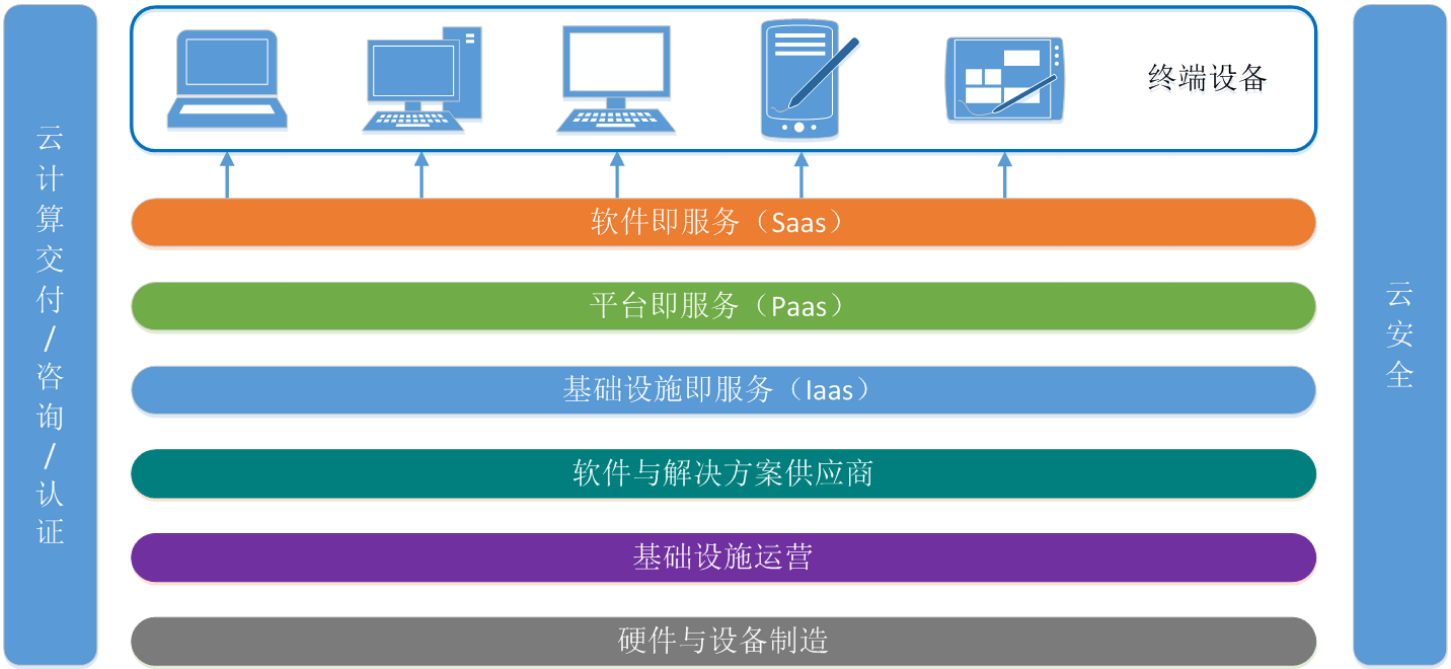
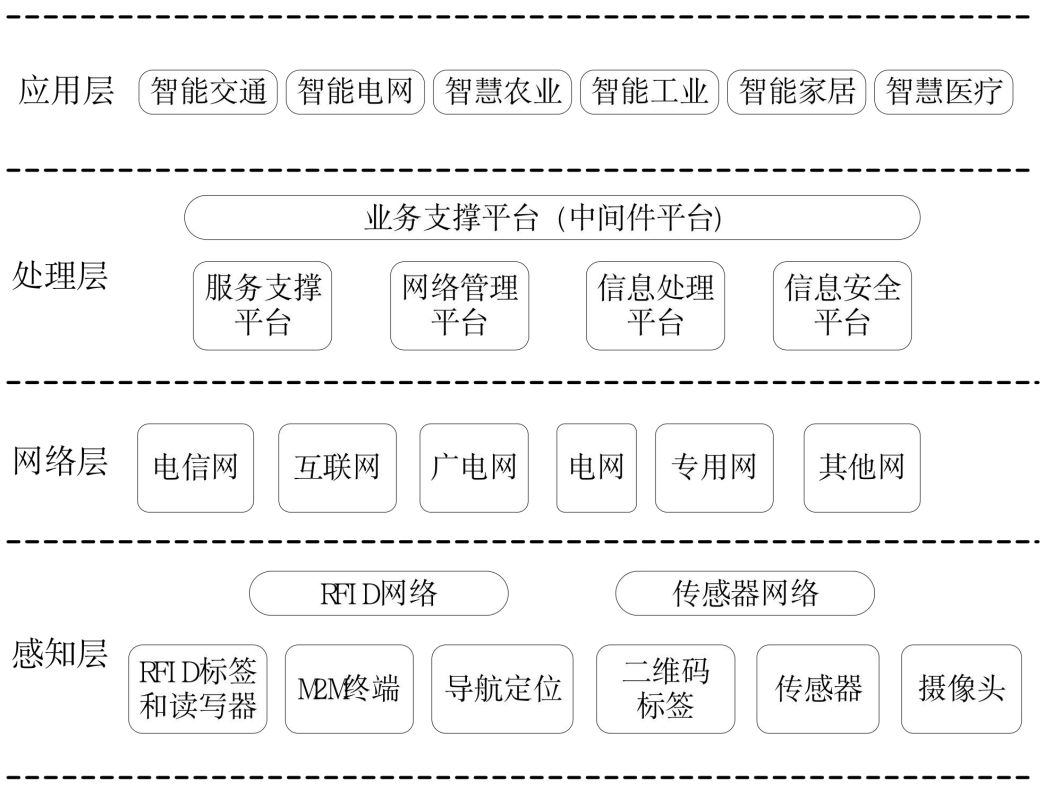


图1-8 云计算产业链

物联网

1. 物联网概念

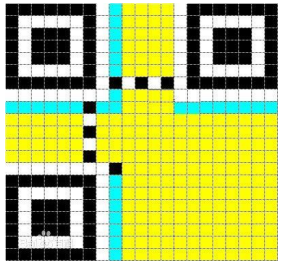
- 物联网是物物相连的互联网，是互联网的延伸，它利用局部网络或互联网等通信技术把传感器、控制器、机器、人员和物等通过新的方式联在一起，形成人与物、物与物相联，实现信息化和远程管理控制



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



矩阵式二维码

采用RFID芯片的公交卡

(a)温湿度传感器

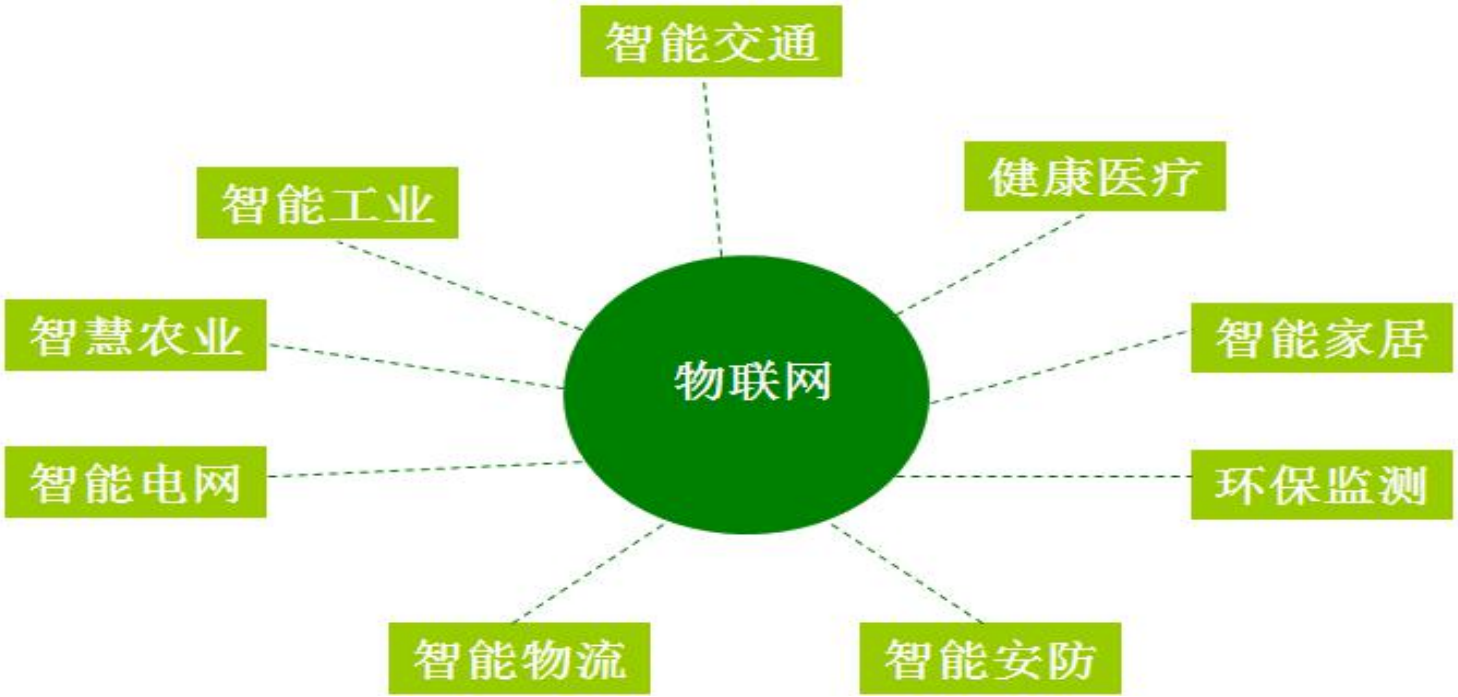
(b)压力传感器

(c)烟雾传感器

物联网

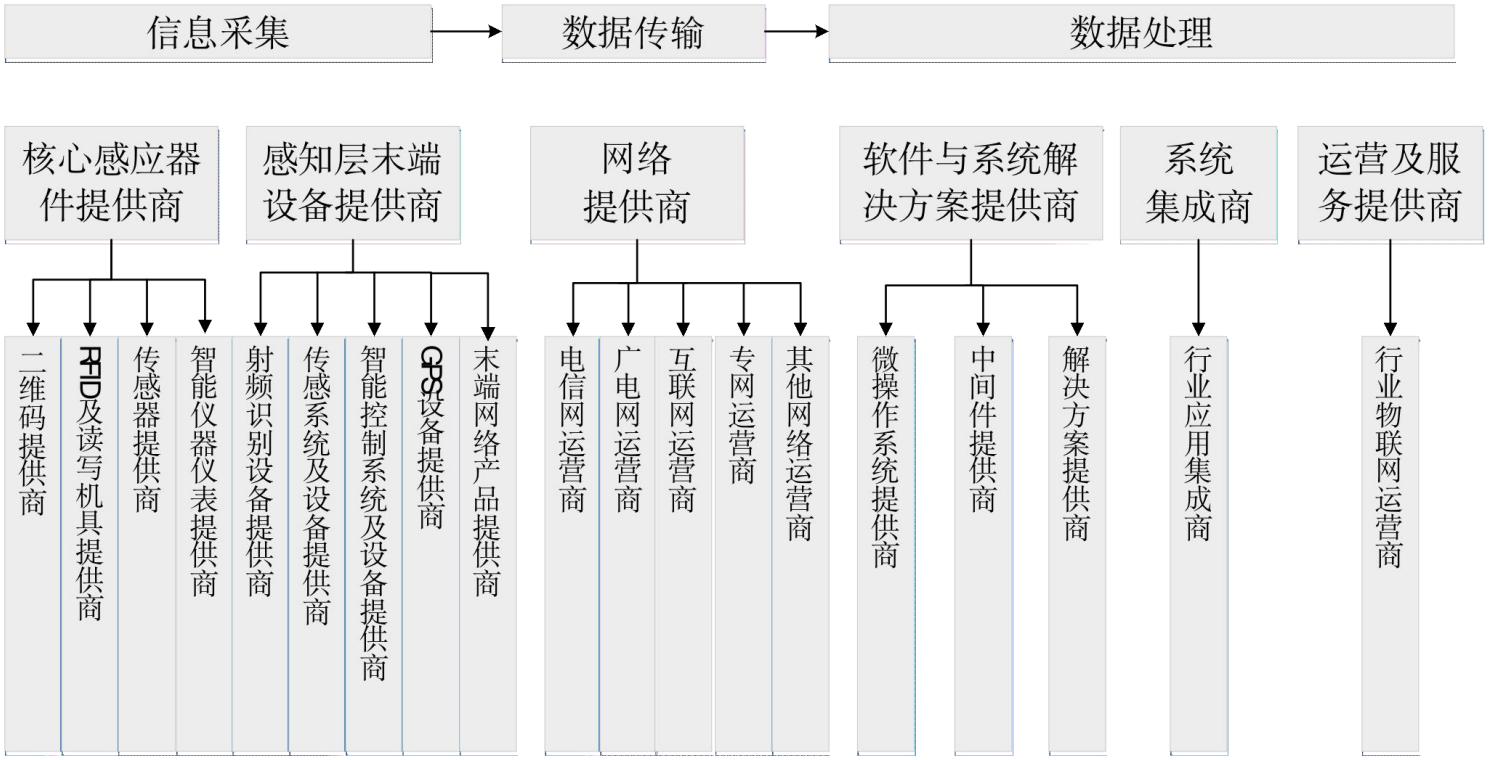
3.物联网应用

- 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



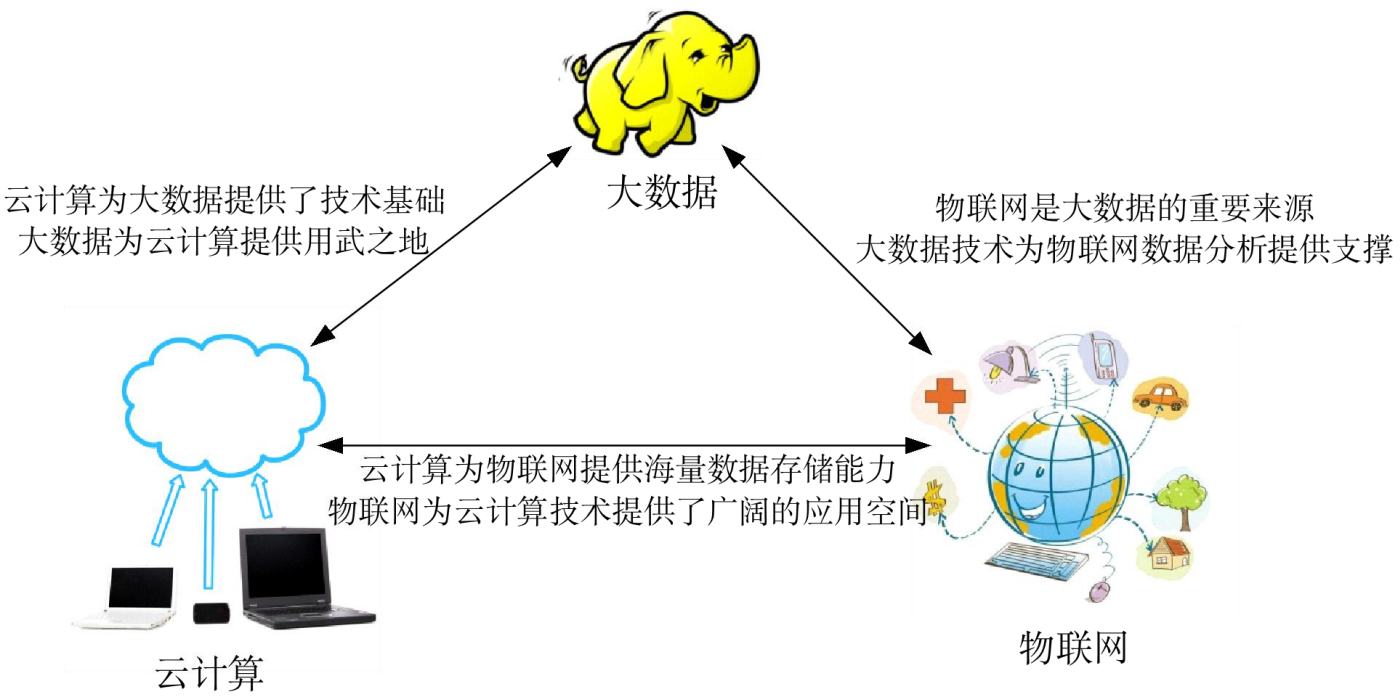
4.物联网产业

- 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



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



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

Thank You!