大数据产业化应用型公开课

云计算、大数据、物联网技术及其产业化应用案例

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

海峡云计算与大数据应用研究中心

E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

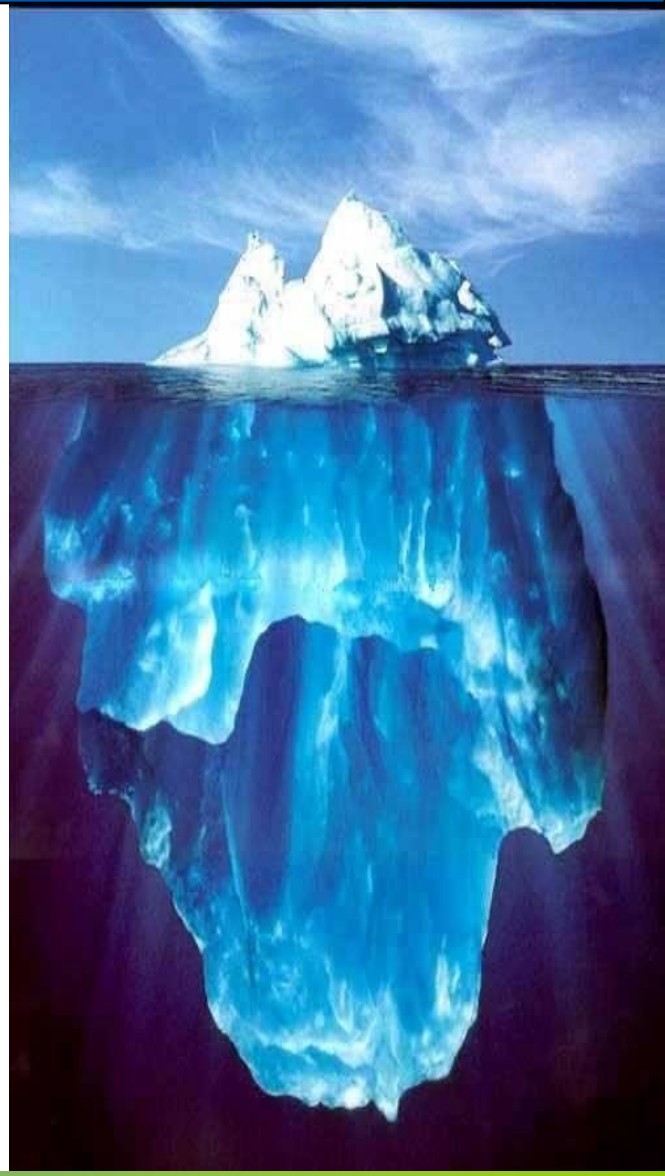
课堂奖品


21世纪高等教育计算机规划教材
COMPUTER
大数据技术原理与应用
——概念、存储、处理、分析与应用
Principles and Applications of Big Data Technology-Big Data Conception, Storage, Processing, Analysis and Application
林子雨 编著


搭建起通向“大数据知识空间”的桥梁和纽带
构建知识体系、阐明基本原理、引导初级实践、了解相关应用
为读者在大数据领域“深耕细作”奠定基础、指明方向
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

提纲

1 大数据时代
2 大数据概念
3 大数据的影响
4 大数据的应用
5 大数据关键技术
6 大数据计算模式
7 大数据理念与实践
8 大数据与云计算、物联网的关系
9产业化应用案例分享



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
GARTNER
MASSIVELY
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

1.2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

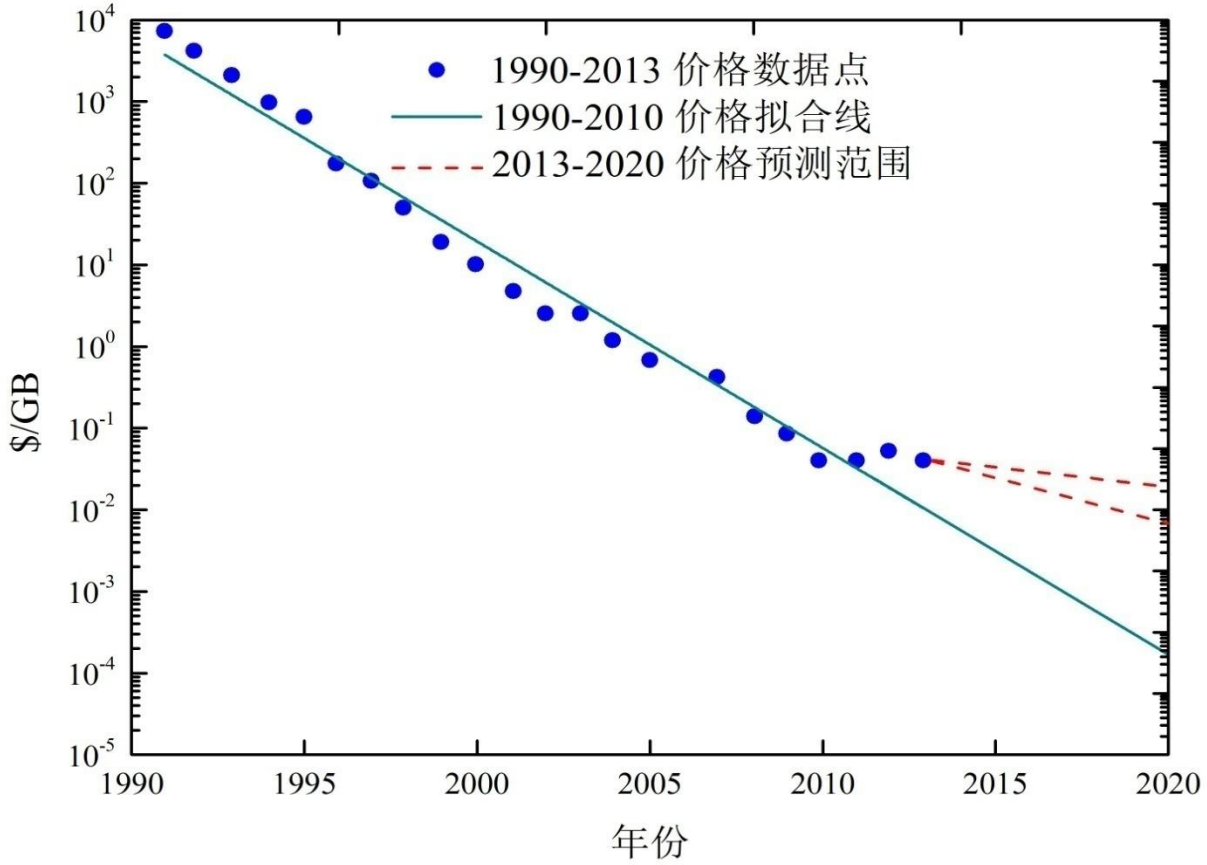


图1-1 存储价格随时间变化情况

1.2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

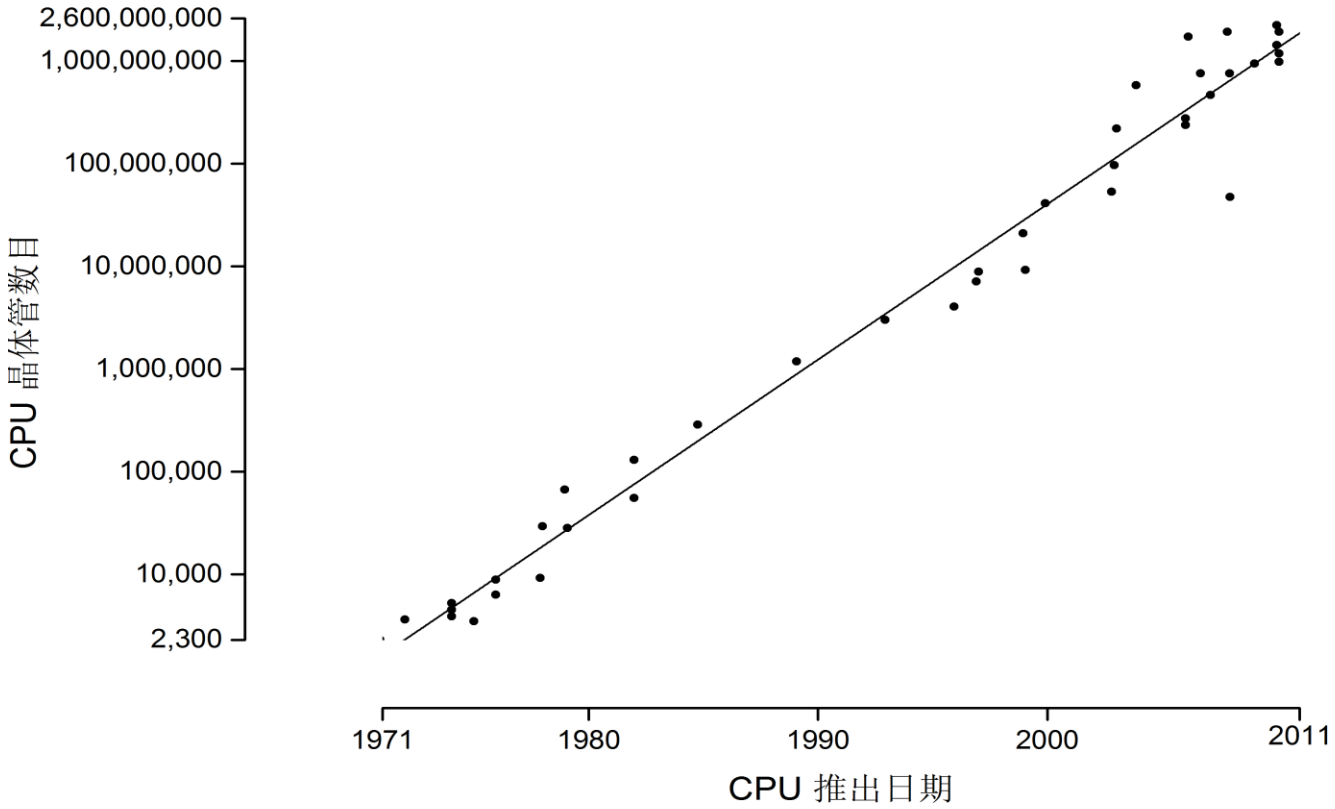


图1-3 CPU晶体管数目随时间变化情况

1.2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

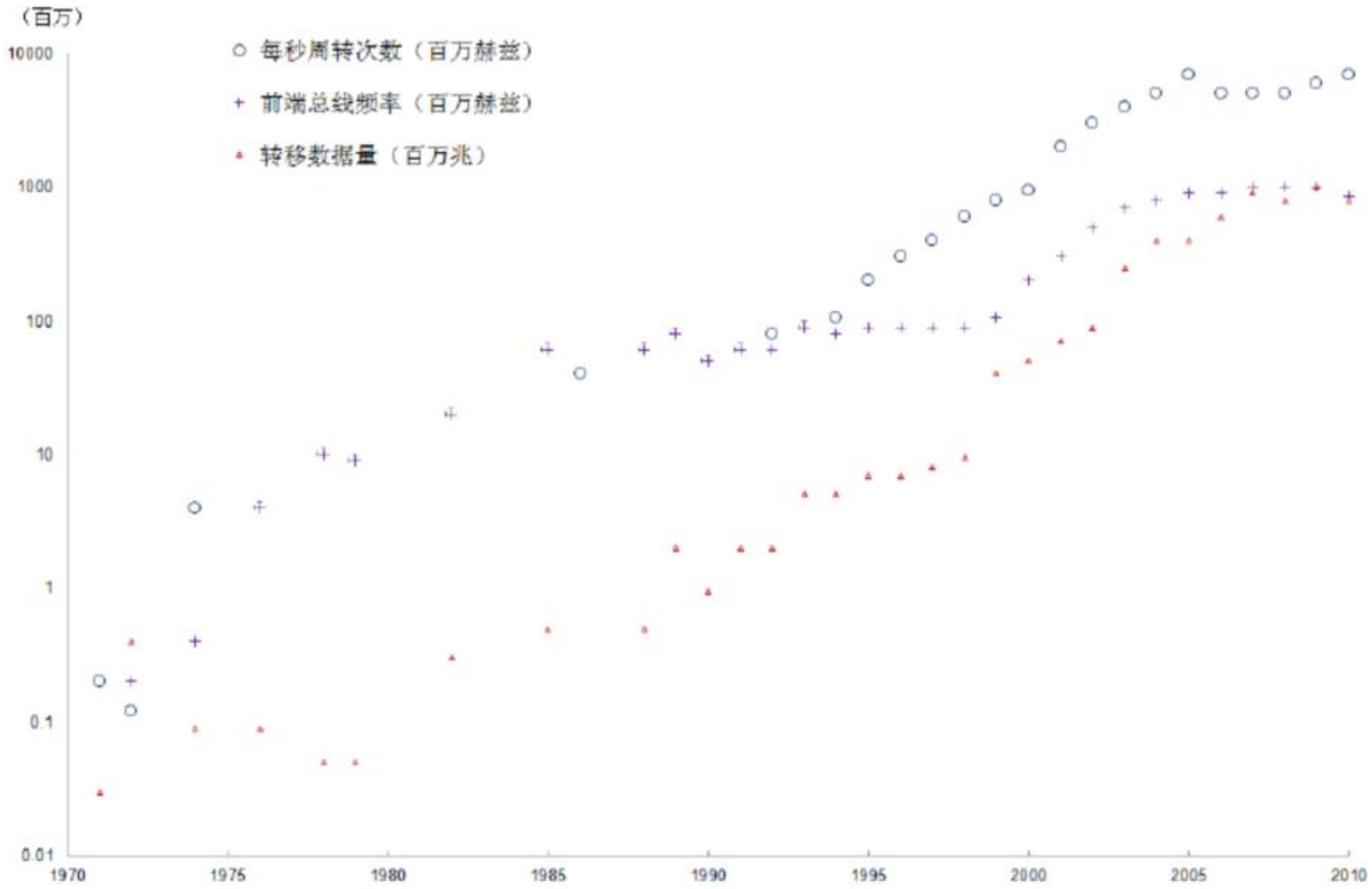


图1-4 网络带宽随时间变化情况

1.3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

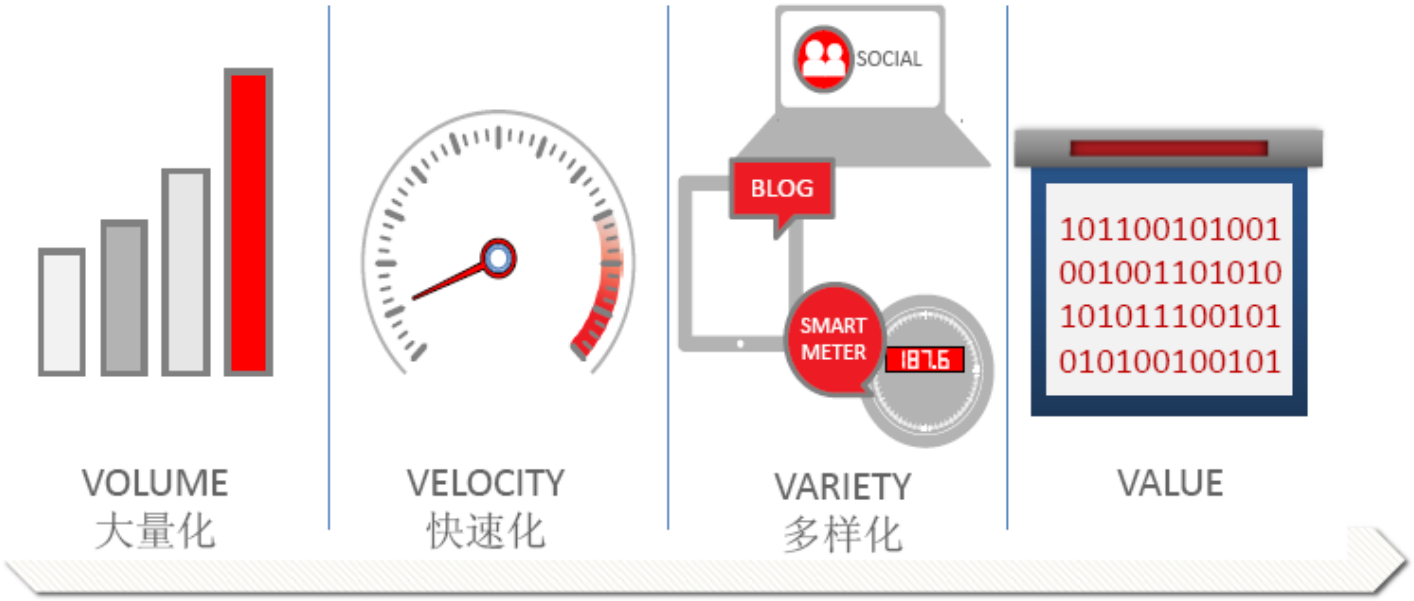
- 数据爆发产生于 Web 2.0 时代，而 Web 2.0 的最重要标志就是用户原创内容
- 智能手机等移动设备加速内容产生
- 数据产生方式是主动的

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

图1-5 数据产生方式的变革

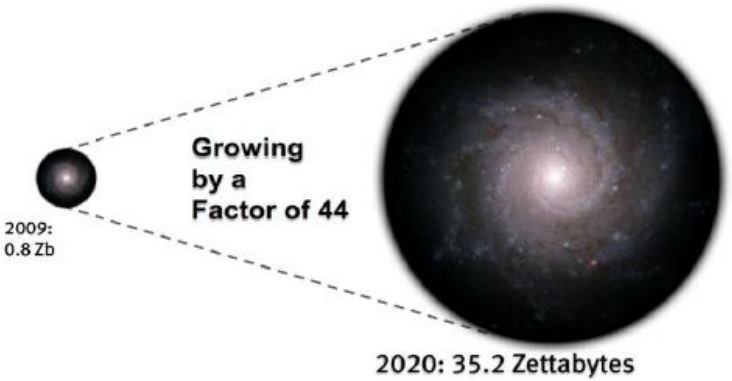
2大数据概念

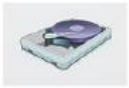


大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

2.1 数据量大

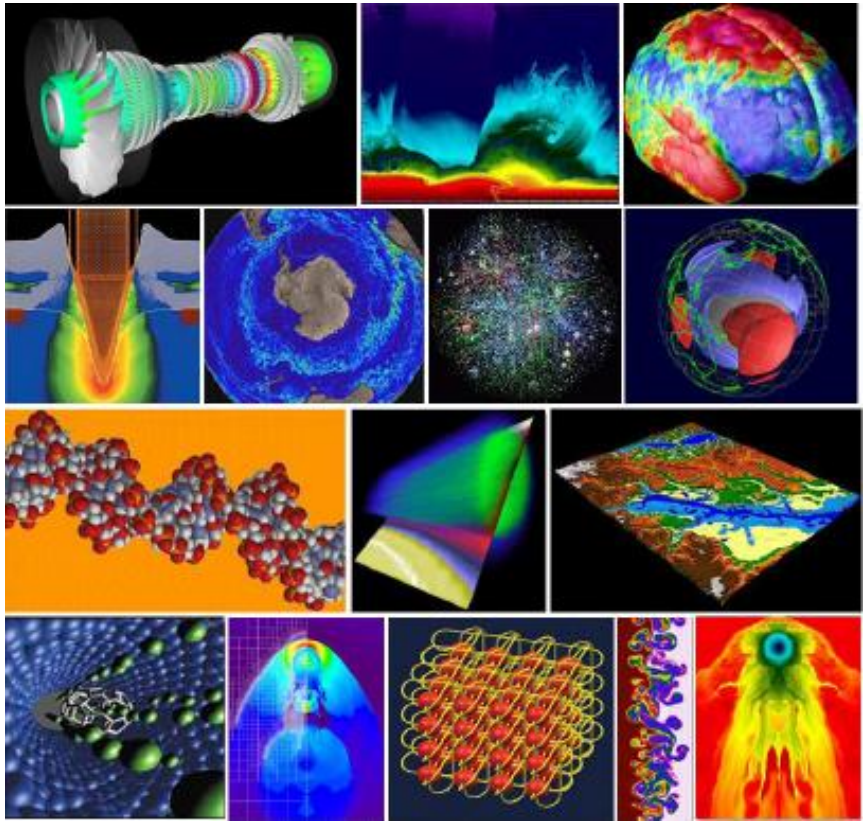
- 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）
- 人类在最近两年产生的数据量相当于之前产生的全部数据量
- 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍



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

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



- 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**

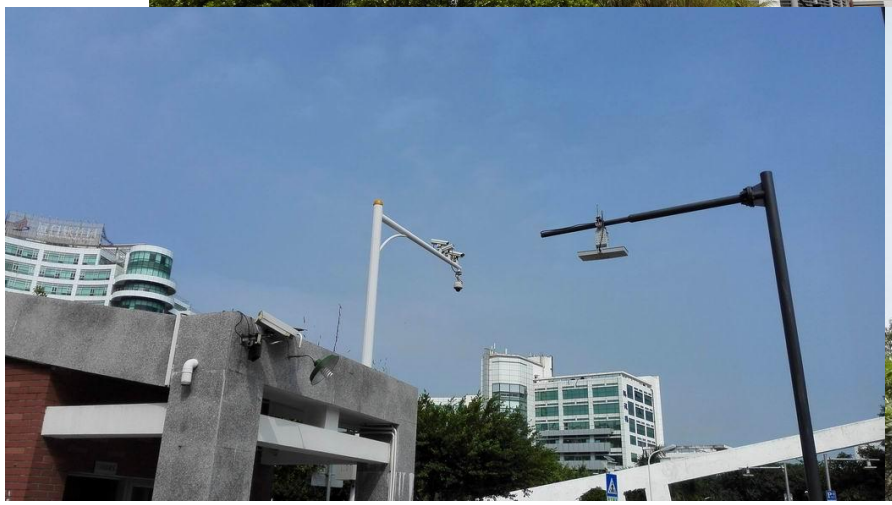
2.3 处理速度快

- 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

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

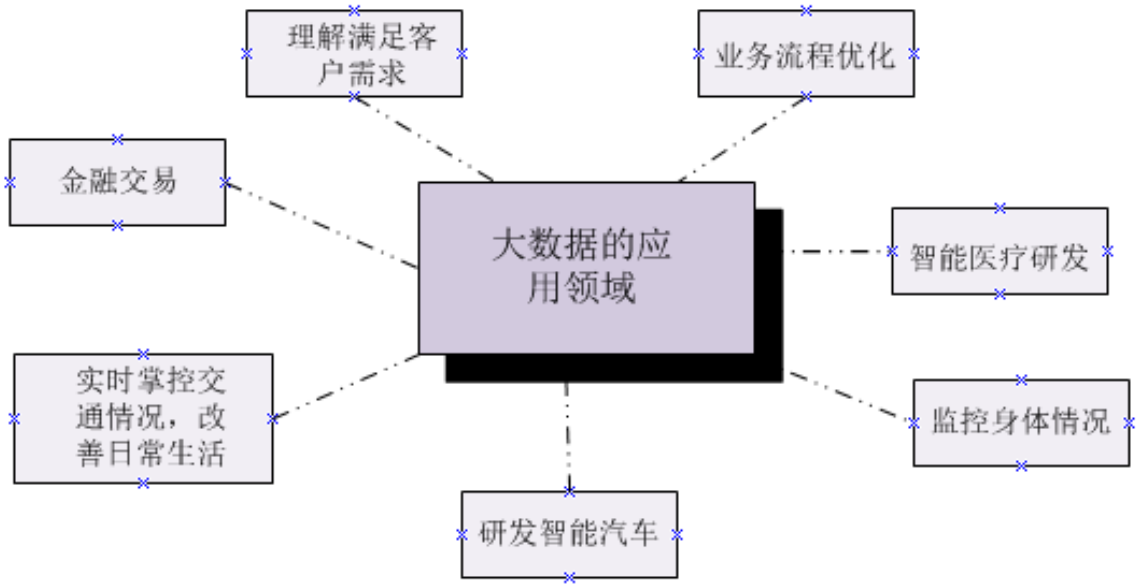


3大数据的影响

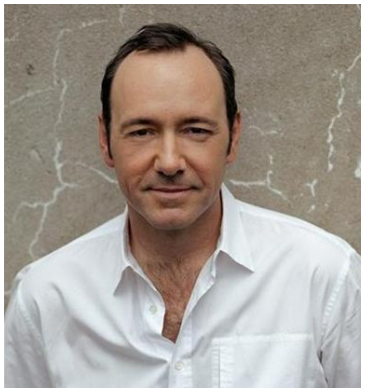
- 大数据对科学研究、思维方式和社会发展都具有重要而深远的影响。
- 在科学研究方面，大数据使得人类科学研究在经历了实验、理论、计算三种范式之后，迎来了第四种范式——数据
- 在思维方式方面，大数据具有“全样而非抽样、效率而非精确、相关而非因果”等三大显著特征，完全颠覆了传统的思维方式
- 在社会发展方面，大数据决策逐渐成为一种新的决策方式，大数据应用有力促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现
- 在就业市场方面，大数据的兴起使得数据科学家成为热门职业
- 在人才培养方面，大数据的兴起，将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制

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

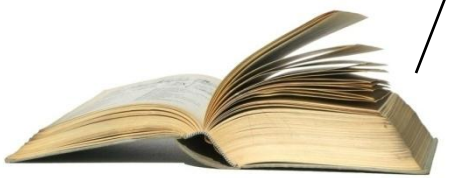


典型的大数据应用实例

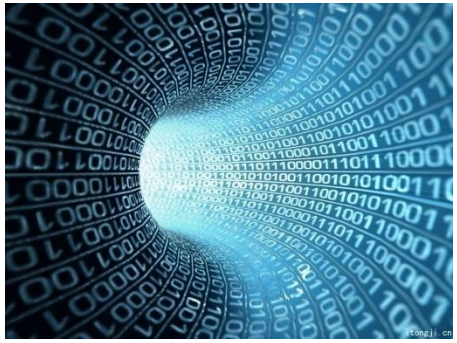


Kevin Spacey

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

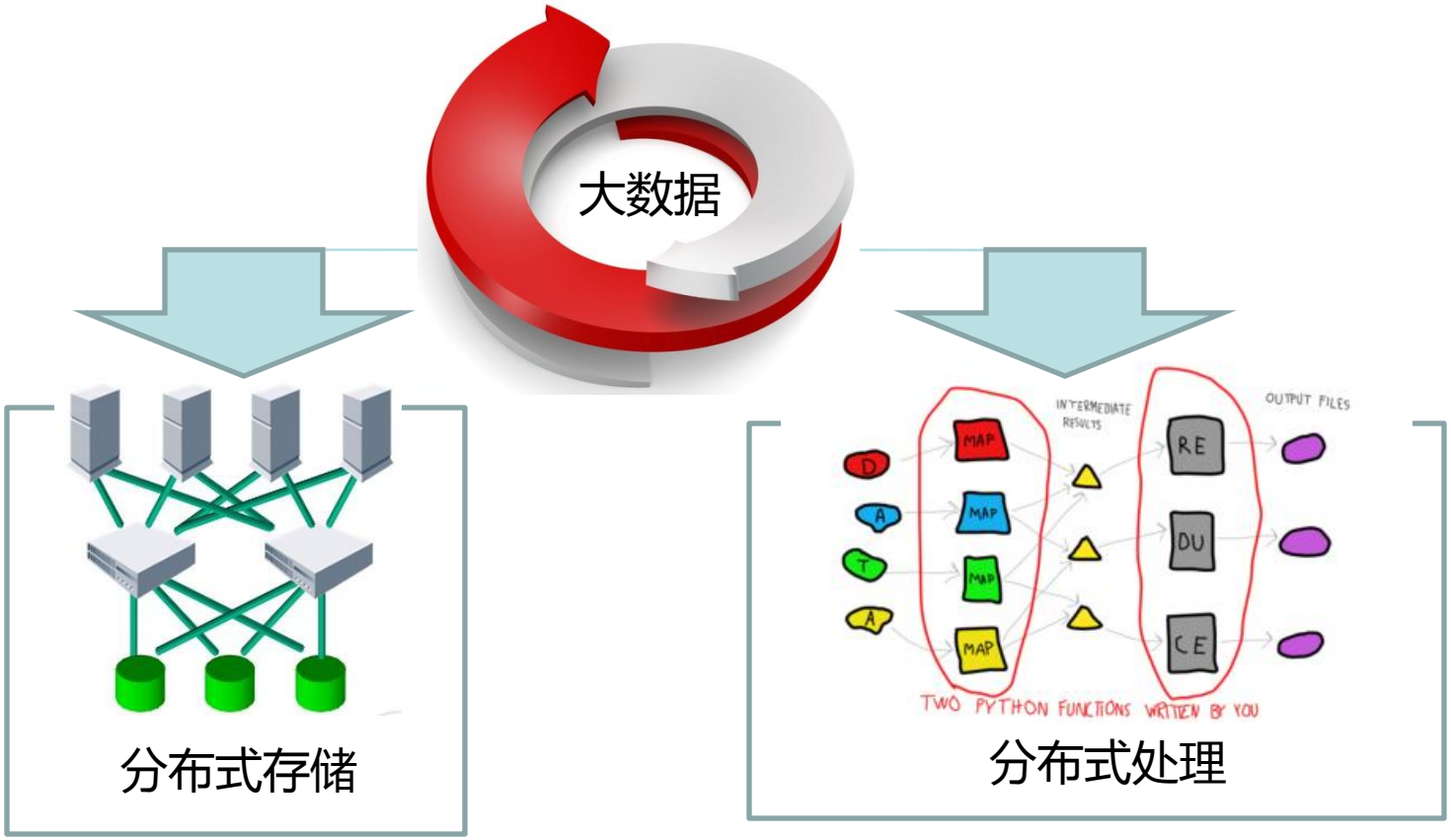
从谷歌流感趋势看大数据的应用价值

“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

典型的大数据应用实例

不是让机器学习逻辑，而是充分发挥机器本身强大的计算能力和数据处理能力

5大数据关键技术



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

MapReduce

5大数据关键技术——Hadoop

- Hadoop是Apache软件基金会旗下的一个开源分布式计算平台，为用户提供了系统底层细节透明的分布式基础架构
- Hadoop是基于Java语言开发的，具有很好的跨平台特性，并且可以部署在廉价的计算机集群中
- Hadoop的核心是分布式文件系统HDFS（Hadoop Distributed File System）和MapReduce
- Hadoop被公认为行业大数据标准开源软件，在分布式环境下提供了海量数据的处理能力
- 几乎所有主流厂商都围绕Hadoop提供开发工具、开源软件、商业化工具和技术服务，如谷歌、雅虎、微软、思科、淘宝等，都支持Hadoop

5大数据关键技术——Hadoop

Hadoop是一个能够对大量数据进行分布式处理的软件框架，并且是以一种可靠、高效、可伸缩的方式进行处理的，它具有以下几个方面的特性：

- 高可靠性
- 高效性
- 高可扩展性
- 高容错性
- 成本低
- 运行在Linux平台上
- 支持多种编程语言

5大数据关键技术——Hadoop

•经过多年的发展，Hadoop项目不断完善和成熟，目前已经包含多个子项目
•除了核心的HDFS和MapReduce以外，Hadoop项目还包括Common、Avro、Zookeeper、HBase、Hive、Chukwa、Pig等子项目，它们提供了互补性服务或在核心层上提供了更高层的服务

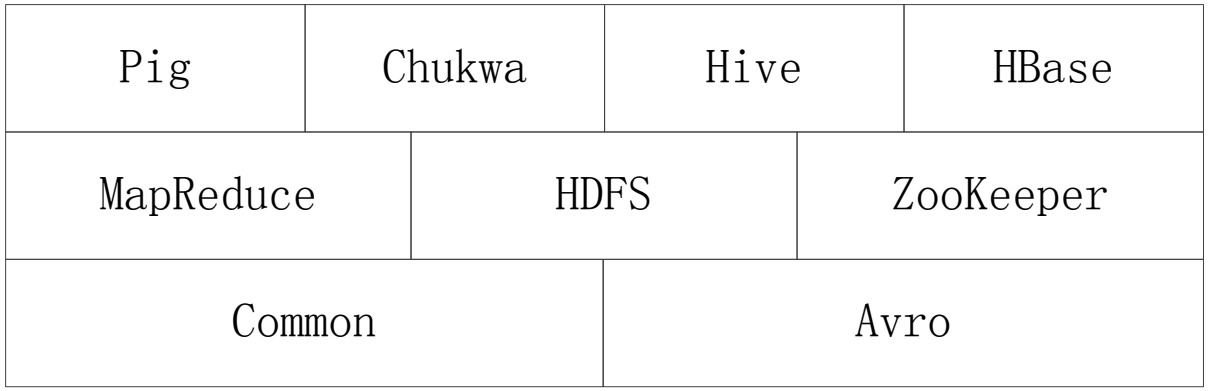


图　Hadoop项目结构图

5大数据关键技术——HDFS

总体而言，HDFS要实现以下目标：

- 兼容廉价的硬件设备
- 流数据读写
- 大数据集
- 简单的文件模型
- 强大的跨平台兼容性

HDFS特殊的设计，在实现上述优良特性的同时，也使得自身具有一些应用局限性，主要包括以下几个方面：

- 不适合低延迟数据访问
- 无法高效存储大量小文件
- 不支持多用户写入及任意修改文件

5大数据关键技术——HDFS

HDFS采用了主从（Master/Slave）结构模型，一个HDFS集群包括一个名称节点（NameNode）和若干个数据节点（DataNode）（如图3-4所示）。名称节点作为中心服务器，负责管理文件系统的命名空间及客户端对文件的访问。集群中的数据节点一般是一个节点运行一个数据节点进程，负责处理文件系统客户端的读/写请求，在名称节点的统一调度下进行数据块的创建、删除和复制等操作。每个数据节点的数据实际上是保存在本地Linux文件系统中的

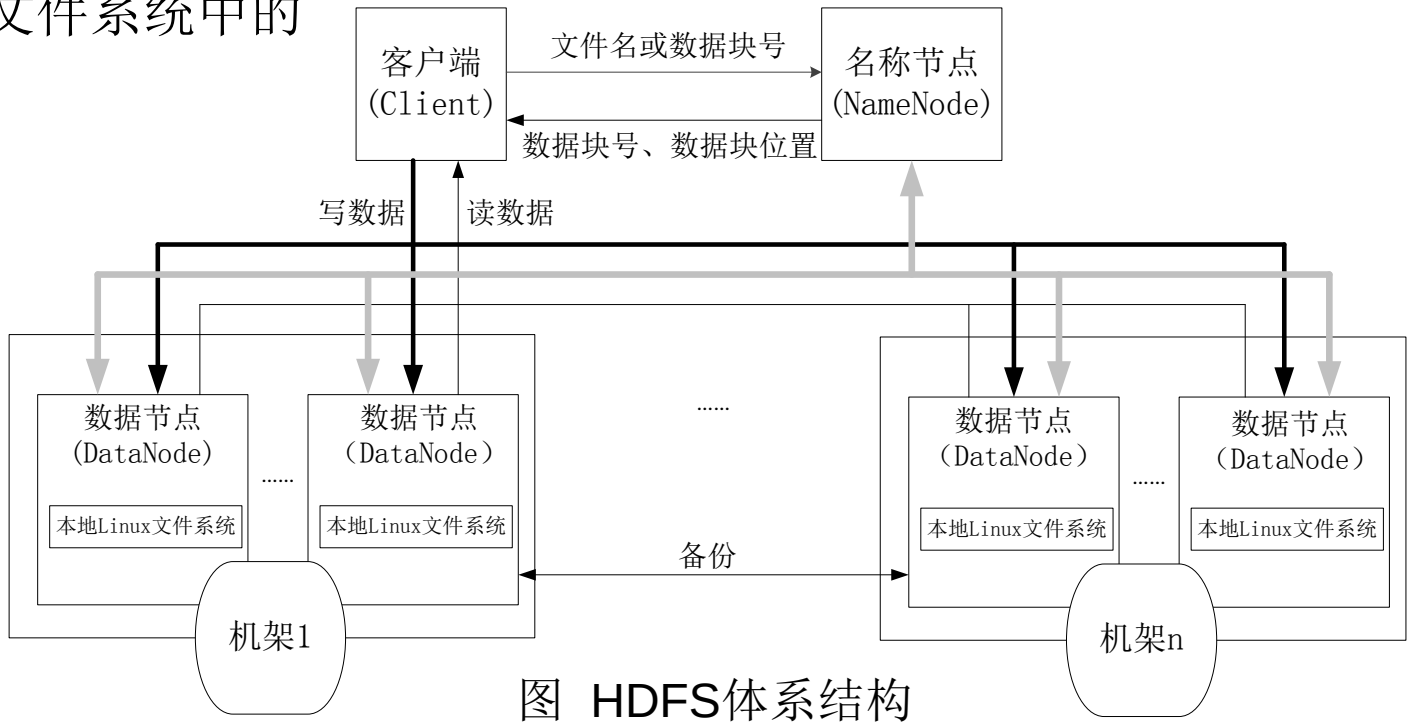


图 HDFS体系结构

5大数据关键技术——MapReduce

- “摩尔定律”，大约每隔18个月性能翻一番
- 从2005年开始摩尔定律逐渐失效，人们开始借助于分布式并行编程来提高程序性能
- 分布式程序运行在大规模计算机集群上，集群中包括大量廉价服务器，可以并行执行大规模数据处理任务，从而获得海量的计算能力
- 谷歌公司最先提出了分布式并行编程模型MapReduce，Hadoop MapReduce是它的开源实现

5大数据关键技术——MapReduce

- MapReduce将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数：Map和Reduce
- 在MapReduce中，一个存储在分布式文件系统中的大规模数据集，会被切分成许多独立的小数据块，这些小数据块可以被多个Map任务并行处理
- MapReduce框架会为每个Map任务输入一个数据子集，Map任务生成的结果会继续作为Reduce任务的输入，最终由Reduce任务输出最后结果，并写入到分布式文件系统中
- MapReduce设计的一个理念就是“计算向数据靠拢”，而不是“数据向计算靠拢”，因为，移动数据需要大量的网络传输开销
- MapReduce框架采用了Master/Slave架构，包括一个Master和若干个Slave。Master上运行JobTracker，Slave上运行TaskTracker
- Hadoop框架是用Java实现的，但是，MapReduce应用程序则不一定要用Java来写

5大数据关键技术——MapReduce

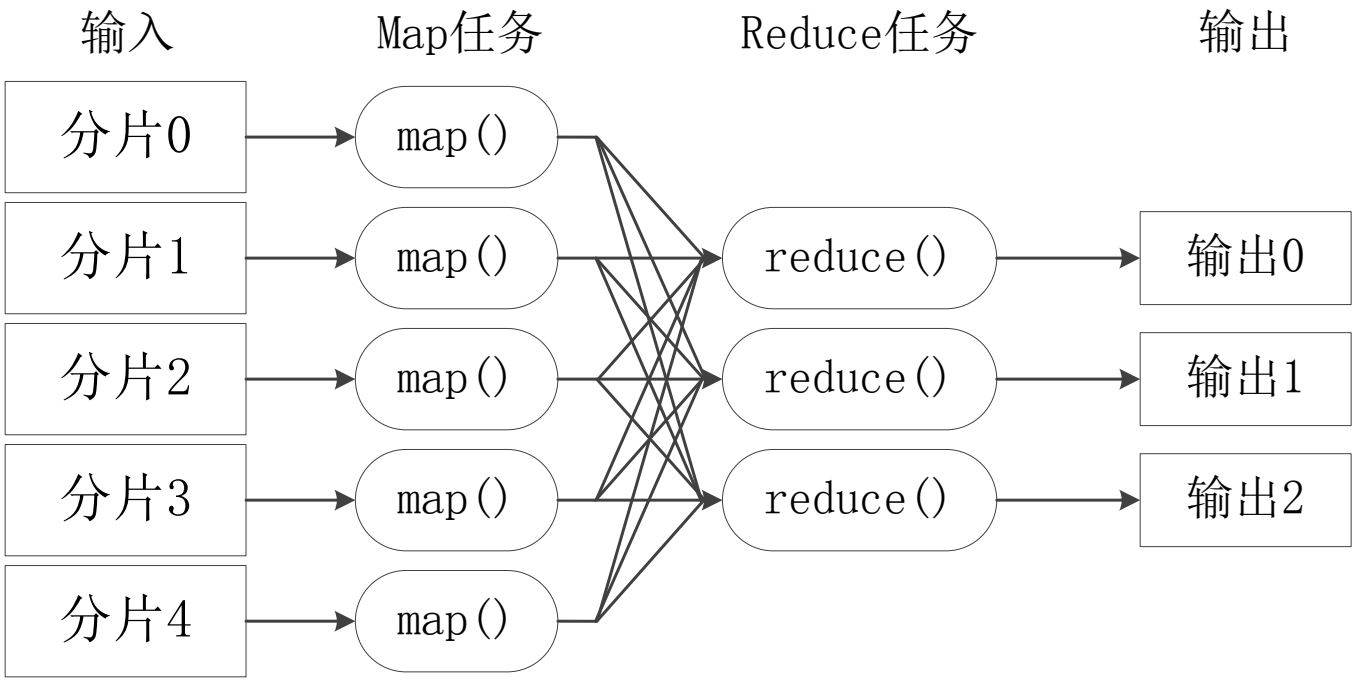


图 MapReduce工作流程

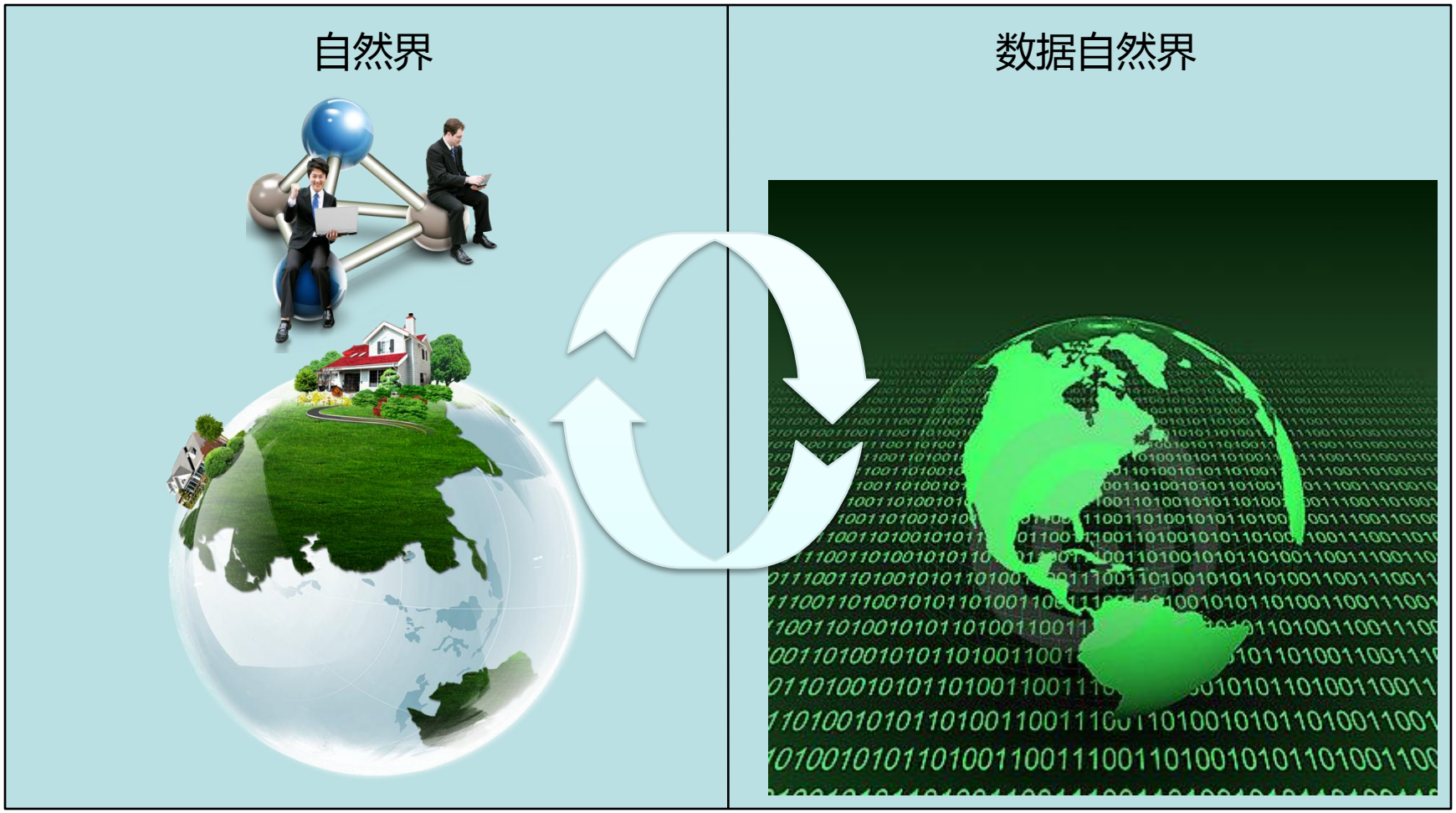
6大数据计算模式

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

7 大数据理念与实践

大数据时代，这是一个以数据为中心的世界……



7 大数据理念与实践

在这个
烟波浩渺的
数据世界里
你的身影
在哪里？

7 大数据理念与实践

在这个数据的世界里，没有个人大数据，就没有你的存在
谁拥有个人大数据，谁就有可能成为这个“浩渺宇宙”中闪亮的星

7 大数据理念与实践

自然界

自然人（林子雨）

数字化

数据自然界

数字教师（林子雨）

7 大数据理念与实践

把个人工作相关信息全部数字化到“数据自然界”

林子雨致力于打造中国高校首个“数字教师”

个人报告、项目课题、合作交流

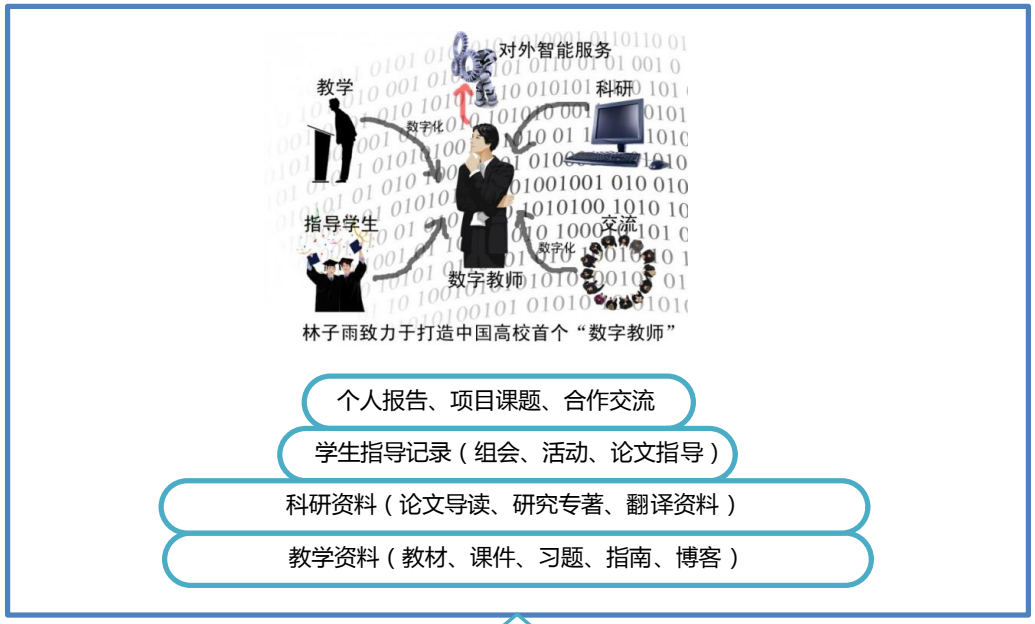
学生指导记录（组会、活动、论文指导）

科研资料（论文导读、研究专著、翻译资料）

教学资料（教材、课件、习题、指南、博客）

7 大数据理念与实践

数据是生产要素，数据产生巨大价值，开放，共享，促进数据自由流通，就是促进社会生产力的发展



教师 服务

教师和其他科研工作者

服务 学生

高校学生和IT公司员工

“数字教师”大平台 形成全面综合的教师个人大数据，提供全方位对外服务

互联网

互联网

7 大数据理念与实践

数字教师的优点：

1	2	3	4	5
勤劳	**全面**	**奉献**	**永生**	**强大**
永远在线，24小时服务	记录教师生涯所有工作历史信息具有档案价值	开放共享，无私奉献	永远存在，生命永恒	能量数倍于自然界的本体

影响力数据：两个“100万”

2009年至今，林子雨数字教师大平台累计免费网络发布超过100万字研究资料

累计网络访问量超过100万次

7 大数据理念与实践

数字教师终极目标：
教师生涯不断完善个人大数据
建成中国最完备的教师个人数字档案
在65岁（75岁？）退休的时候
入藏国家档案馆
申请国家级数字文化遗产（如果有这个……）

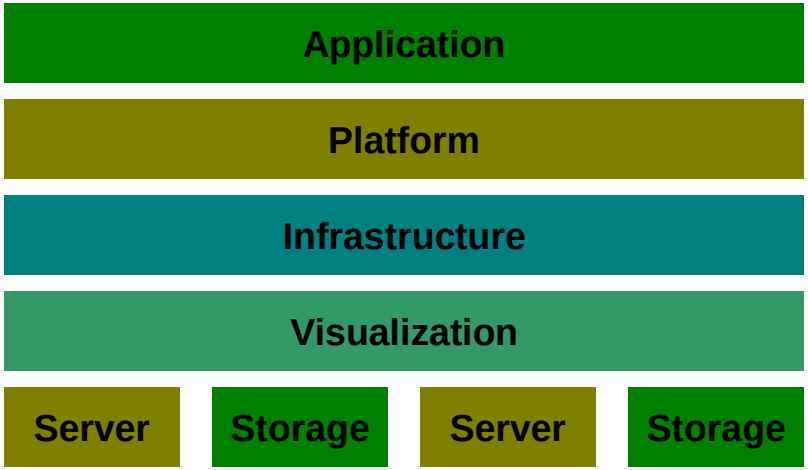
8大数据与云计算、物联网的关系

- 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

8.1云计算

1. 云计算概念

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

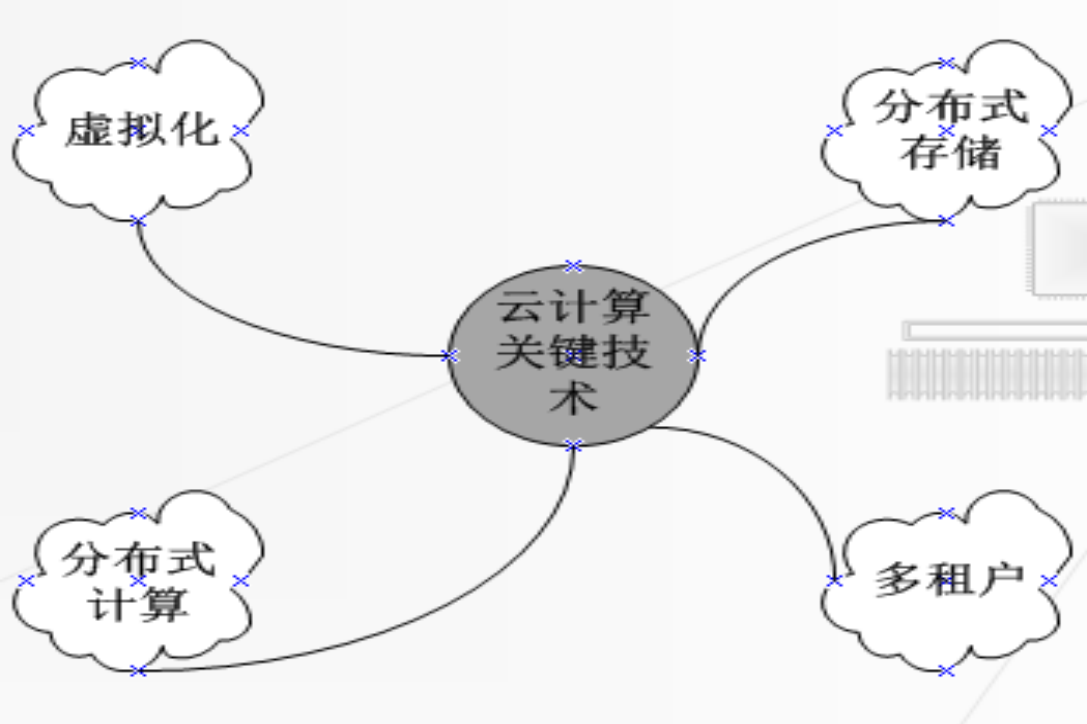


SaaS	Software as a Service	Google Apps, Microsoft “Software+Services”
PaaS	Platform as a Service	IBM IT factory, Google App Engine, Force.com
IaaS	Infrastructure as a Service	Amazon EC2, IBM Blue Cloud, Sun Grid
dSaaS	data Storage as a Service	Nirvanix SDN, Amazon S3, Cleversafe dsNet

8.1云计算

2. 云计算关键技术

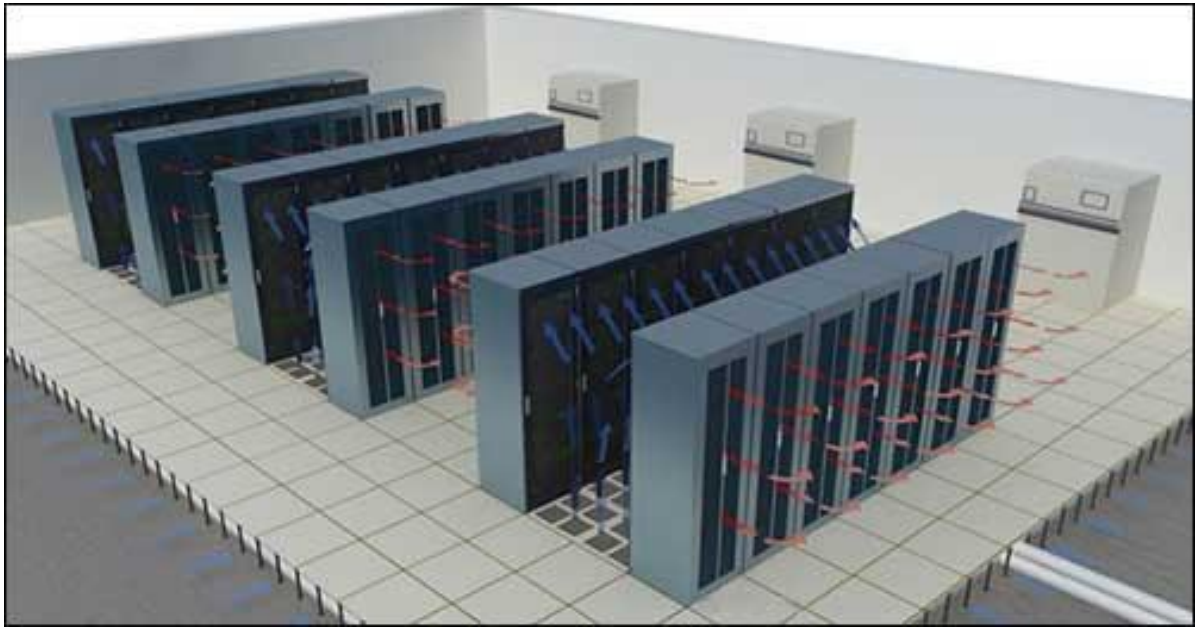
- 云计算关键技术包括：虚拟化、分布式存储、分布式计算、多租户等



8.1云计算

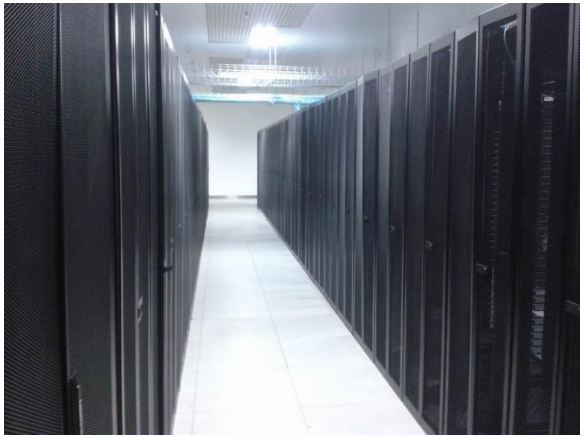
3. 云计算数据中心

- 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



中国国际信息技术（福建）产业园

国国际信息技术（福建）产业园



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

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

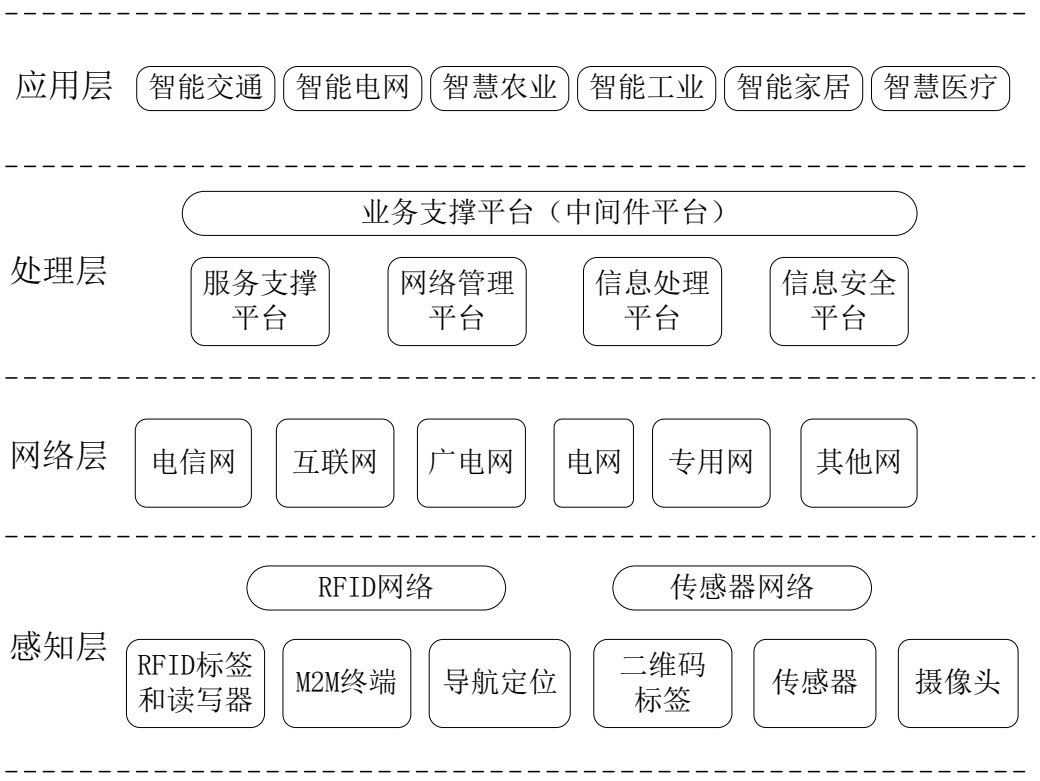


图1-9 物联网体系架构

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

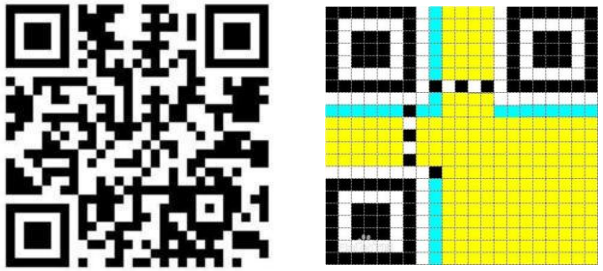


图1-10 矩阵式二维码

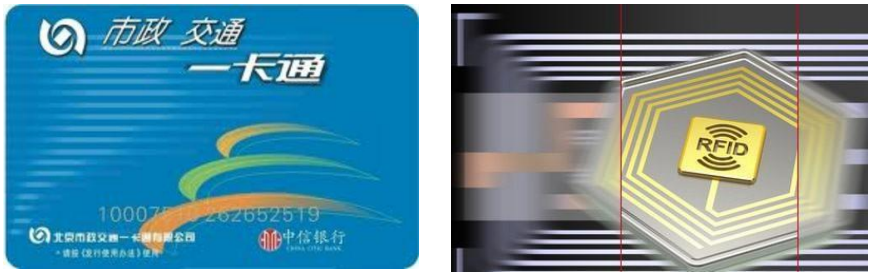


图1-11 采用RFID芯片的公交卡

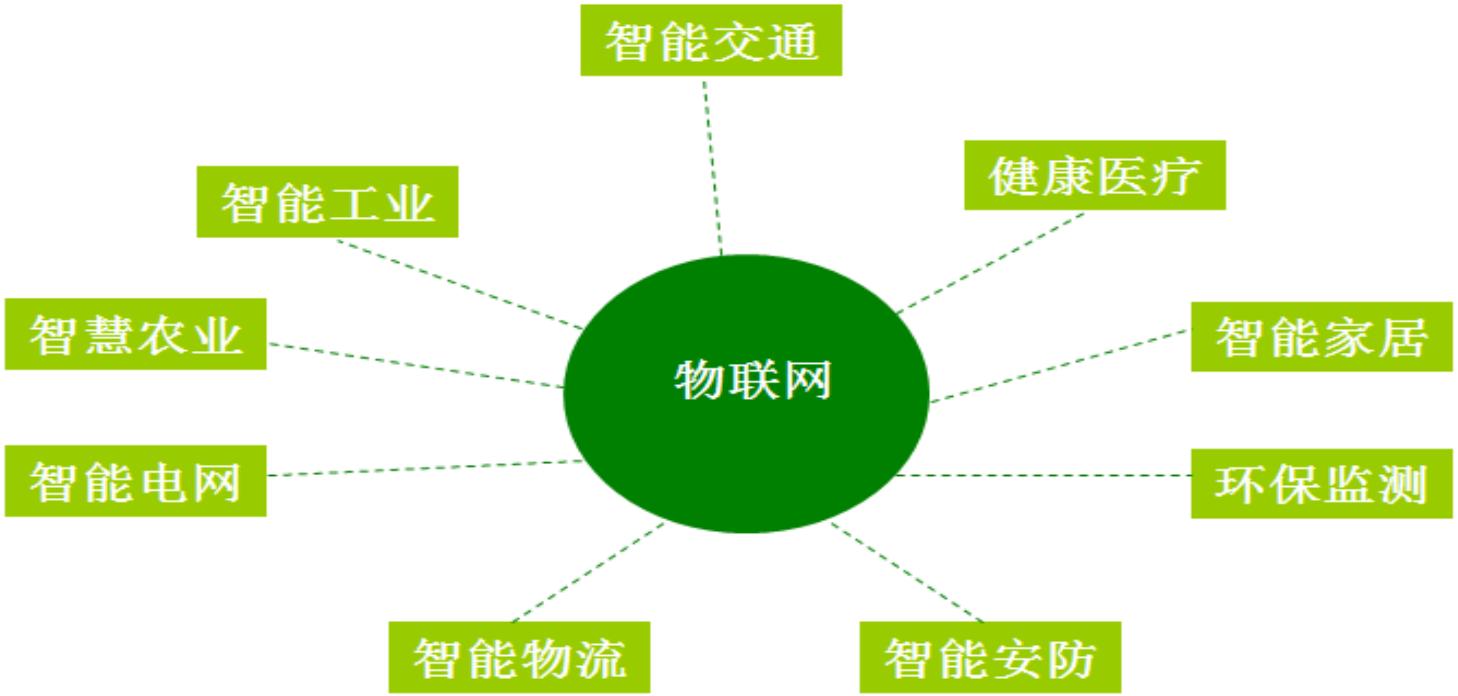
(a)温湿度传感器　(b)压力传感器　(c)烟雾传感器

图1-12 不同类型的传感器

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



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

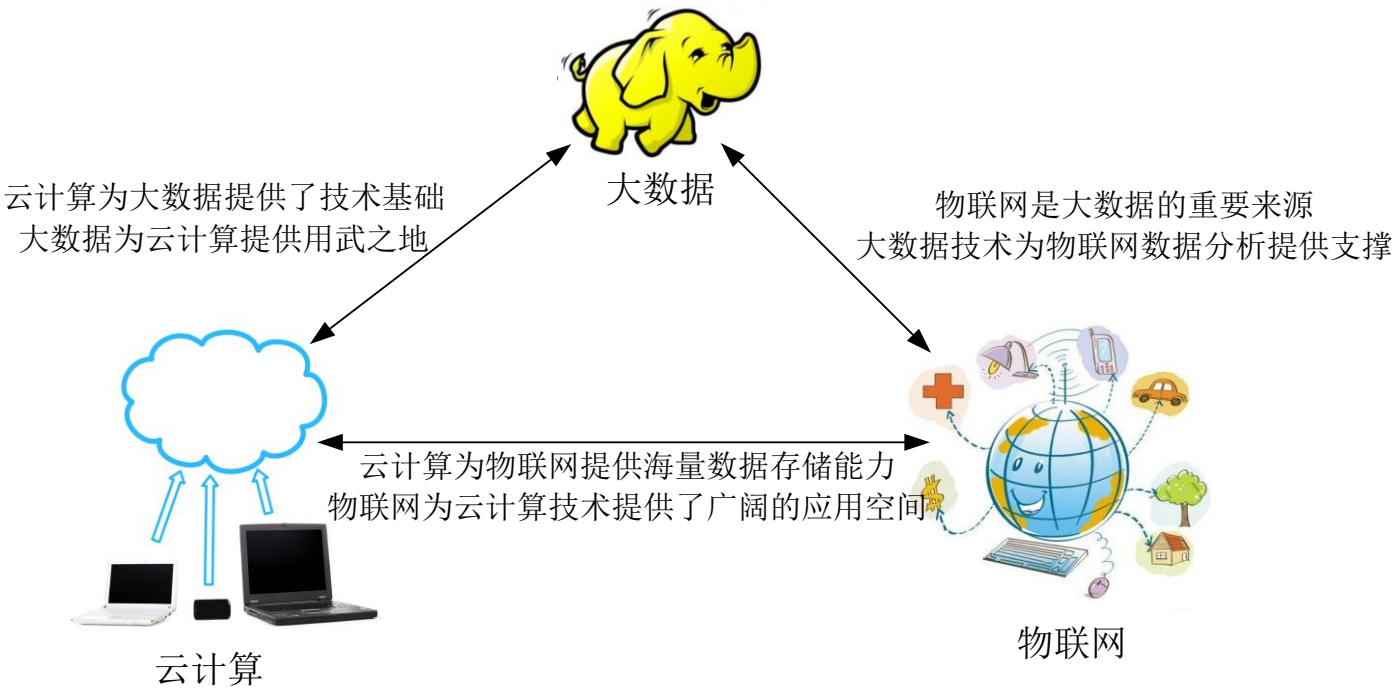


图1-9 大数据、云计算和物联网之间的关系

9产业化应用案例分享

案例一：在物流行业的应用

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

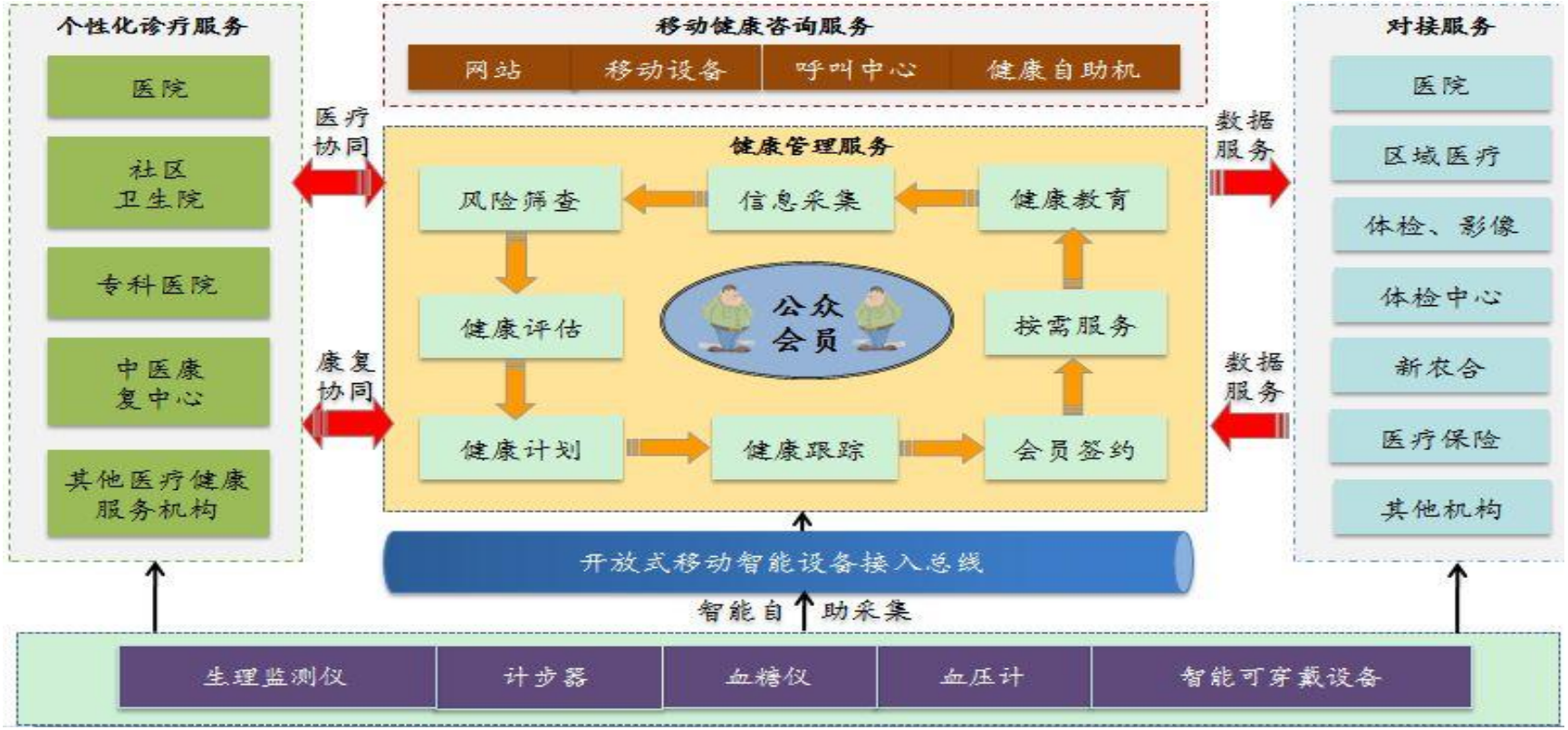
关键举措四：构建开放数据应用平台，向物流生态系统内各种群提供服务

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

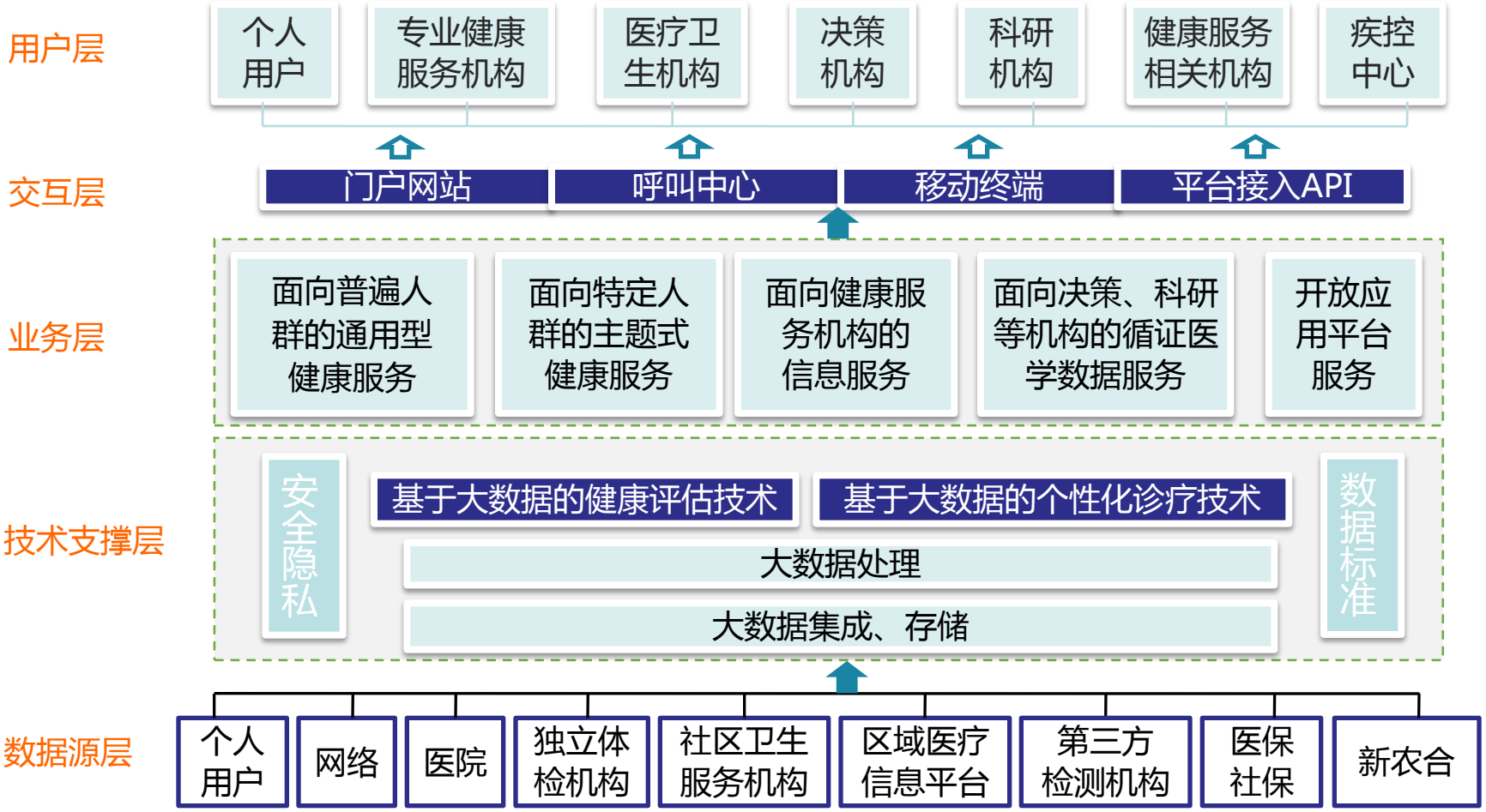
案例二：在医疗健康行业的应用

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



案例三：在餐饮配送行业的应用

云配送

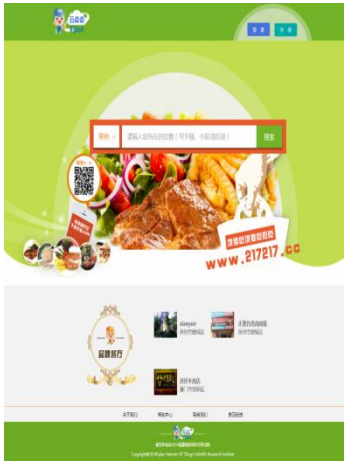
“云配送”系统是一款基于云计算技术的在线软件系统，以微信平台系统为技术支撑，微信用户为目标消费群体，服务于全国线下实体商家及线下网络商家的微信营销系统产品。

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

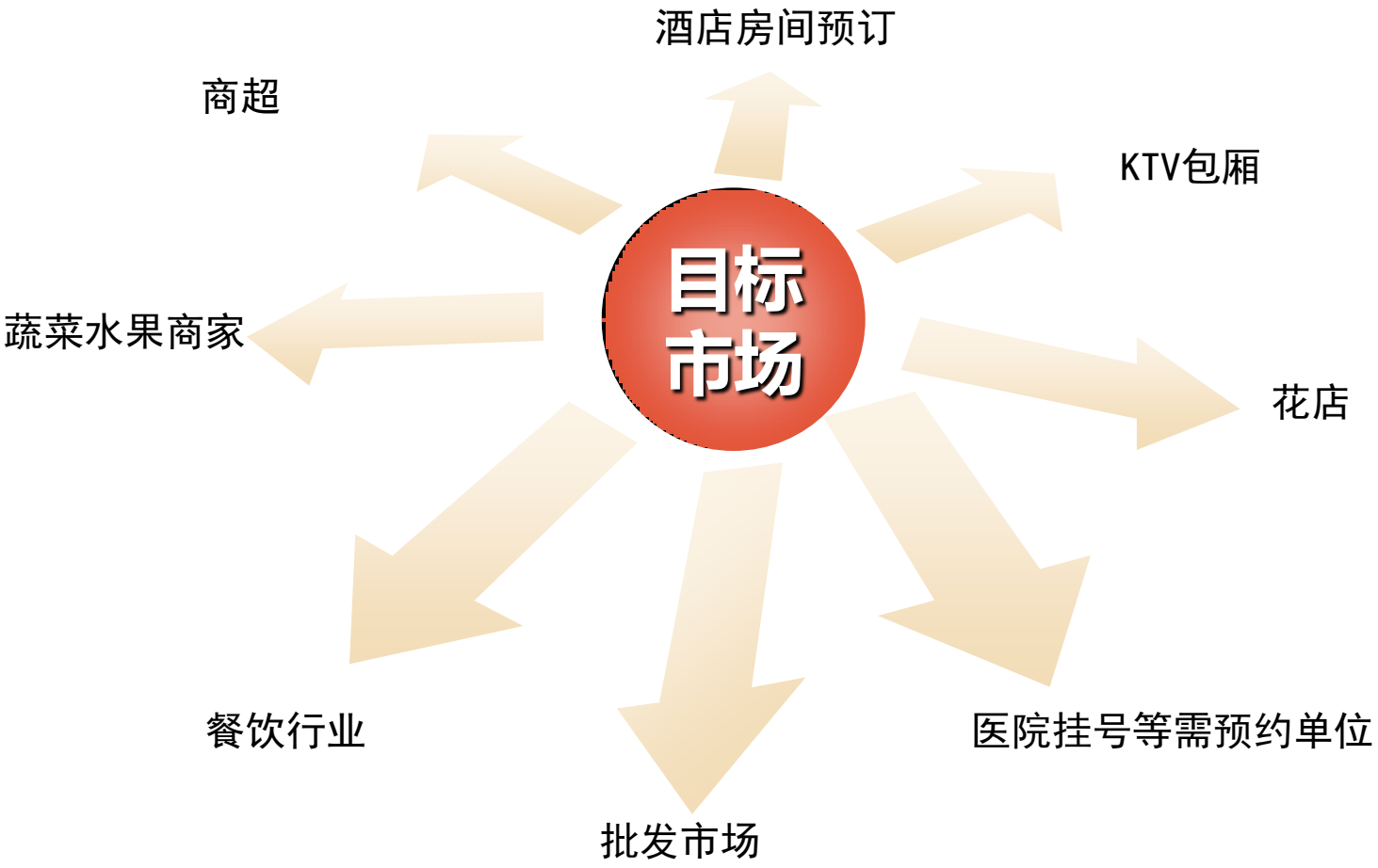
微信下单

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

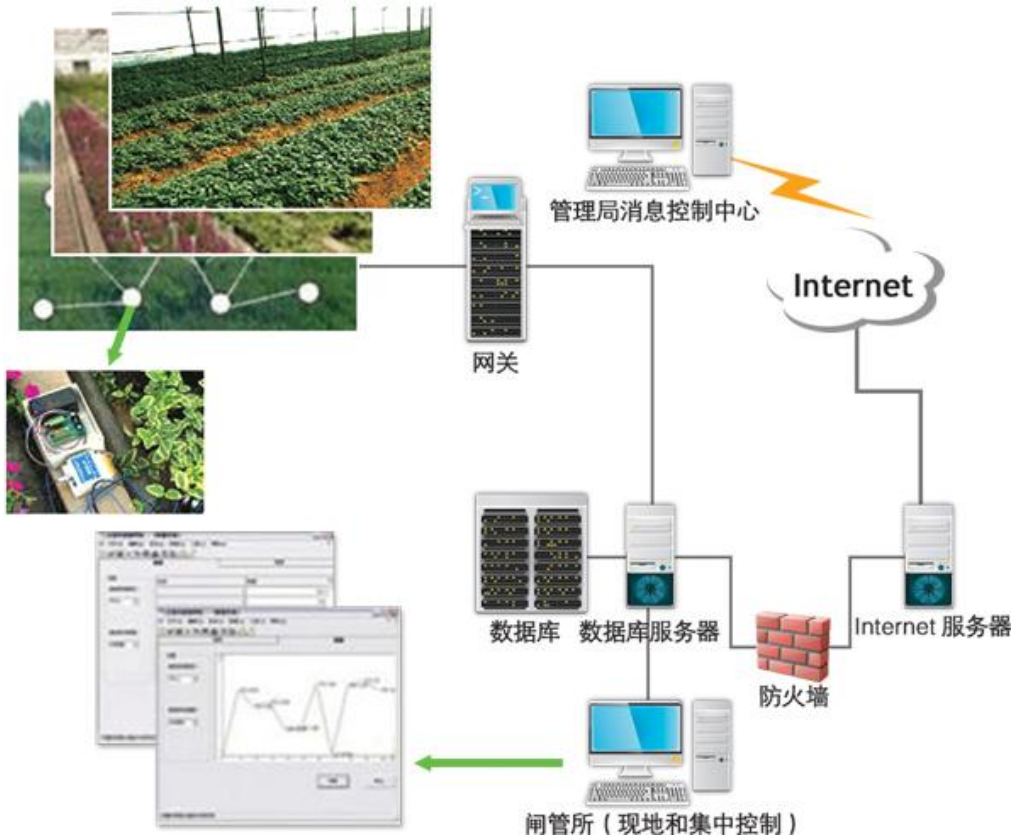


案例四：在菜篮子工程中的应用

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

小结

- 简要介绍了云计算、大数据、物联网概念及其相互关系
- 呈现行业案例：智能物流、综合健康服务平台
- 分享了一点心得体会

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