



# 大数据技术原理与应用

刘磊  
浙江财经大学数据科学学院



浙江财经大学  
Zhejiang University of Finance & Economic

数据  $\overset{?}{\neq}$  信息

“数据是爆炸了，信息却很贫乏”

数据是反映客观事物属性的记录，是信息的具体表现形式；  
数据经过加工处理之后成为信息。

# 学习内容

数据采集



kafka



数据存储



数据处理与分析



数据可视化

.....



Scala



python

编程语言

# 01

## 大数据概述

- 大数据时代的背景
- 大数据的定义和特征
- 大数据的发展历程
- 大数据技术
- 大数据应用
- 云计算，物联网和大数据的关系



# 数据度量

---



1Byte

1KB = 1,024 Bytes

1MB = 1,024 KB

1GB = 1,024 MB

1TB = 1,024 GB

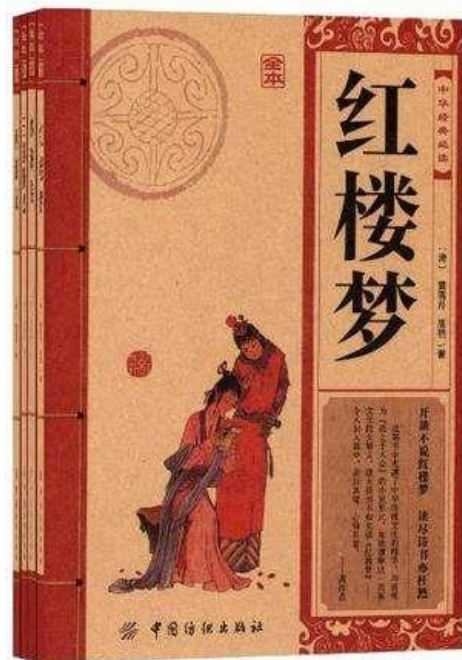
1PB = 1,024 TB

1EB = 1,024 PB

1ZB = 1,024 EB

1YB = 1,024 ZB

# 数据度量



《红楼梦》含标点87万字（不含标点853509字）  
每个汉字占两个字节：

1汉字=2bytes

1GB 约等于 671部红楼梦

1TB 约等于 631,903 部

1PB 约等于 647,068,911部

美国国会图书馆藏书（151,785,778册）  
（2011年4月：收录数据235TB）

# 大数据时代的背景

21世纪是数据信息大发展的时代，移动互联网、物联网、社交网络、电子商务等极大拓展了互联网的边界和应用范围，各种数据正在迅速膨胀并变大。

互联网（搜索、电商）、移动互联网（微博、微信）、物联网（传感器）、GPS、医学影像、安全监控、金融（银行、股市、保险）、电信（通话、短信）都在疯狂产生着数据。



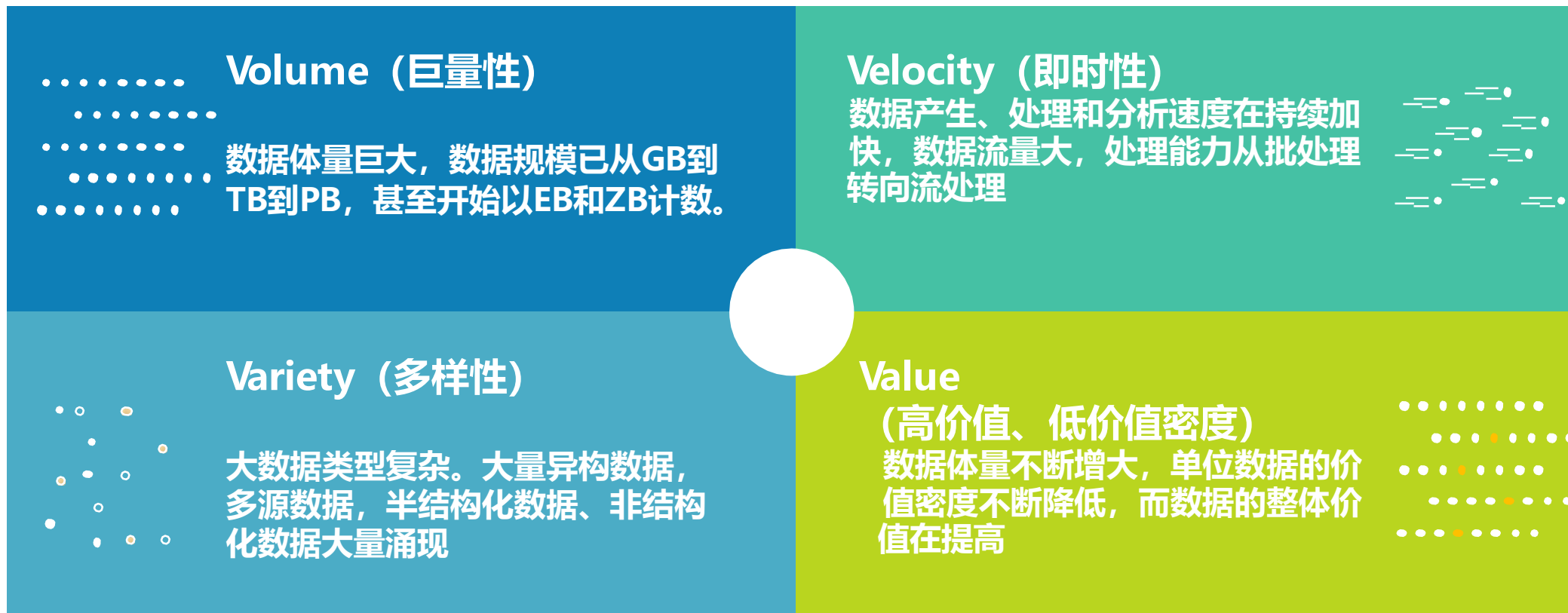
# 大数据的定义

- **麦肯锡**：大数据是指大小**超出常规**的数据库工具获取、存储、管理和分析能力的数据集。  
(并不是说一定要超过特定TB的数据集才能算大数据)
- **Gartner公司**：大数据是**需要新处理模式**才能具有更强的决策力、洞察发现力和流程优化的**海量、高增长率和多样化的**信息资产。
- **美国国家标准技术研究院 (NIST)**：数据**量大、获取速度快或形态多样**的数据，难以用传统关系型数据分析方法进行有效分析，或者需要大规模的水平扩展才能高效处理。
- **国际数据公司 (IDC)**：从大数据的4个特征来定义，即海量的数据规模 (**Volume**)、数据处理的快速性 (**Velocity**)、多样的数据类型 (**Variety**)、数据价值密度低 (**Value**)，即所谓的4V特性。  
IBM认为大数据还应该具有其**真实性 (Veracity)**。





# 大数据的特征



# 数据量大



1Byte

1KB = 1,024 Bytes

1MB = 1,024 KB

1GB = 1,024 MB

1TB = 1,024 GB

1PB = 1,024 TB

1EB = 1,024 PB

1ZB = 1,024 EB

1YB = 1,024 ZB

一般情况下，大数据是以PB、EB、ZB为单位进行计量的

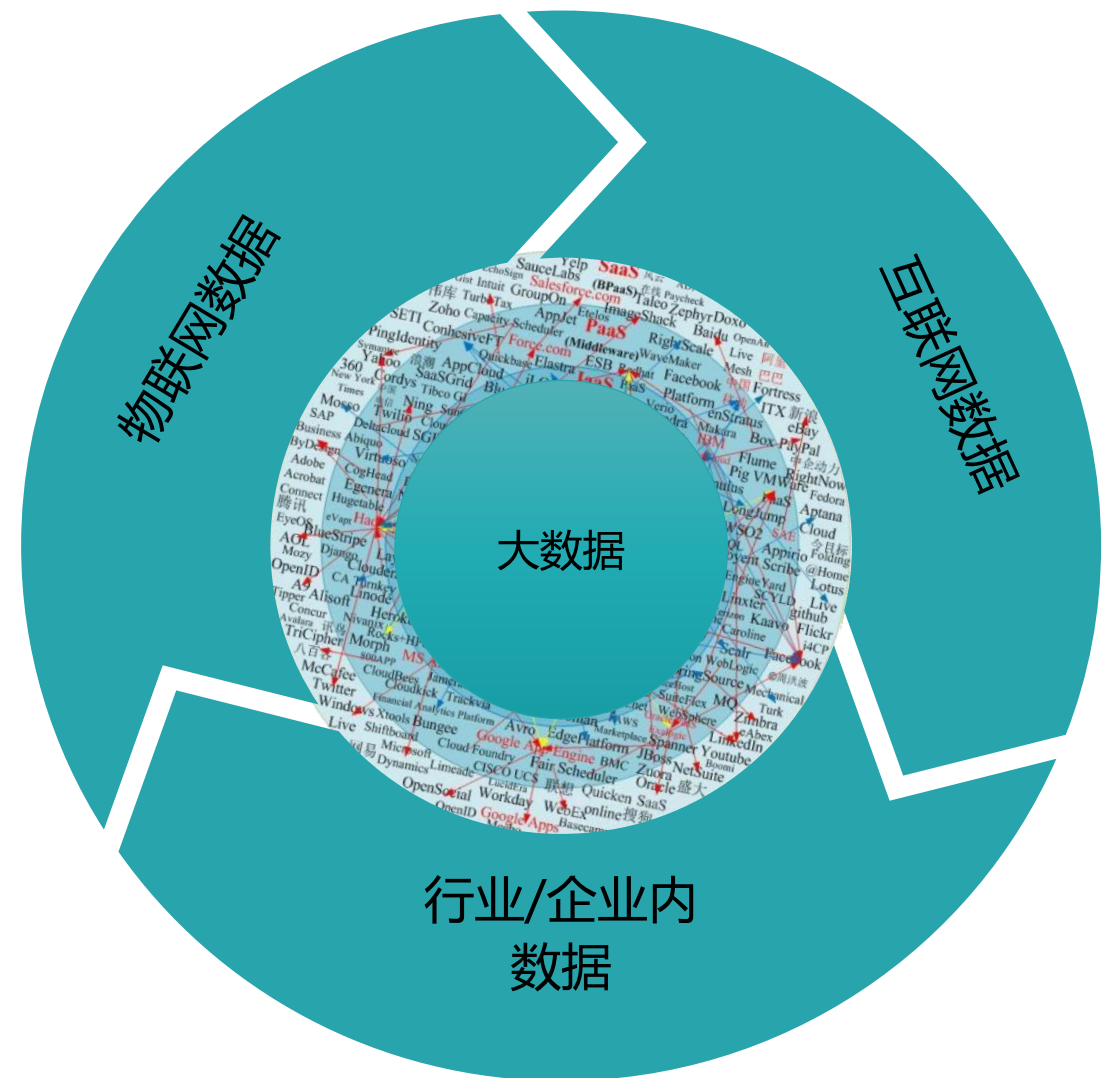
1PB相当于50%的全美学术研究图书馆藏书信息内容

5EB相当于至今全世界人类所讲过的话语

1ZB如同全世界海滩上的沙子数量总和

1YB相当于7000个人体内的微细胞总和

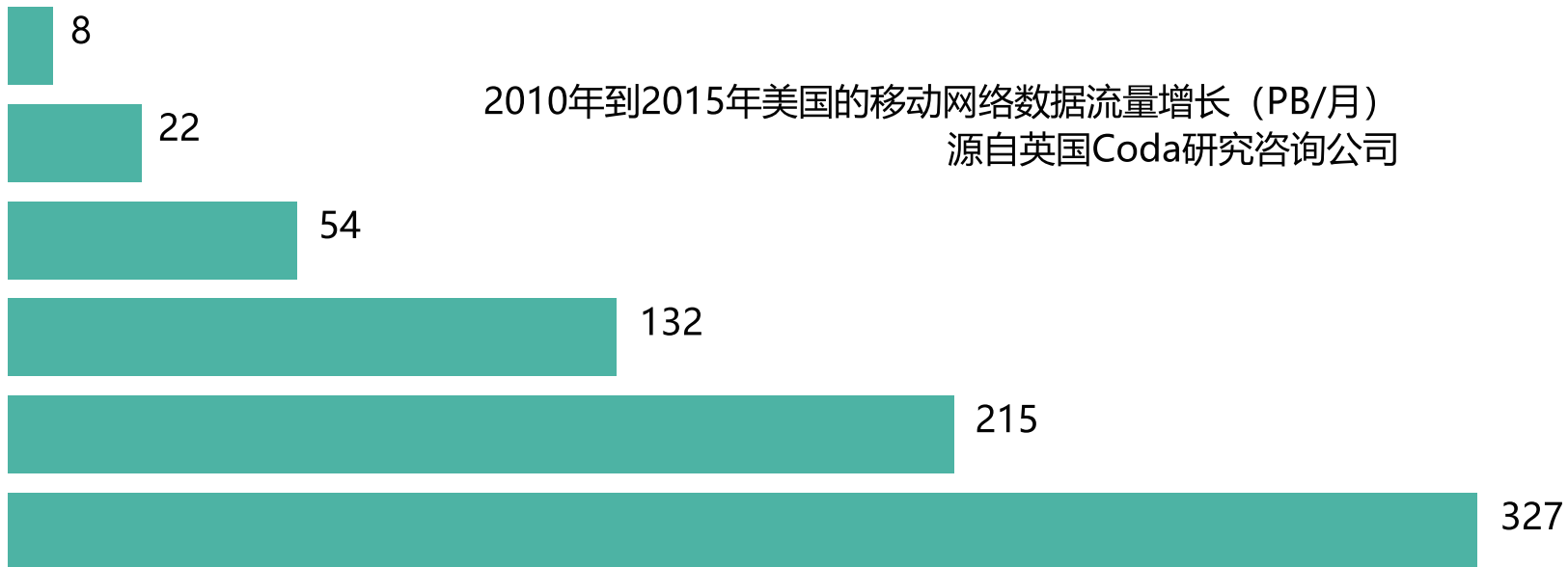
# 数据类型繁多



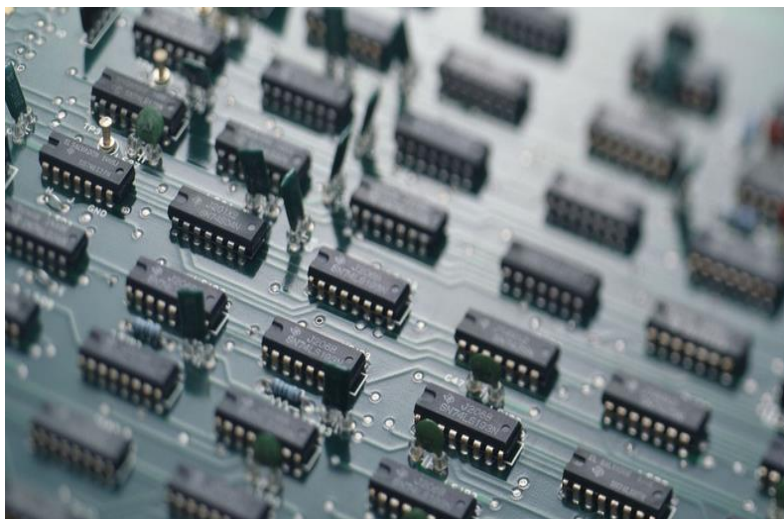
数据来源多	企业内部多个应用系统的数据、互联网和物联网的兴起，带来了微博、社交网站、传感器等多种来源。
数据类型多	保存在关系数据库中的结构化数据只占少数，70~80%的数据是如图片、音频、视频、模型、连接信息、文档等非结构化和半结构化数据。
关联性强	数据之间频繁交互，比如游客在旅行途中上传的图片和日志，就与游客的位置、行程等信息有了很强的关联性。

# 处理速度快

大数据的增长速度快



大数据的处理速度快

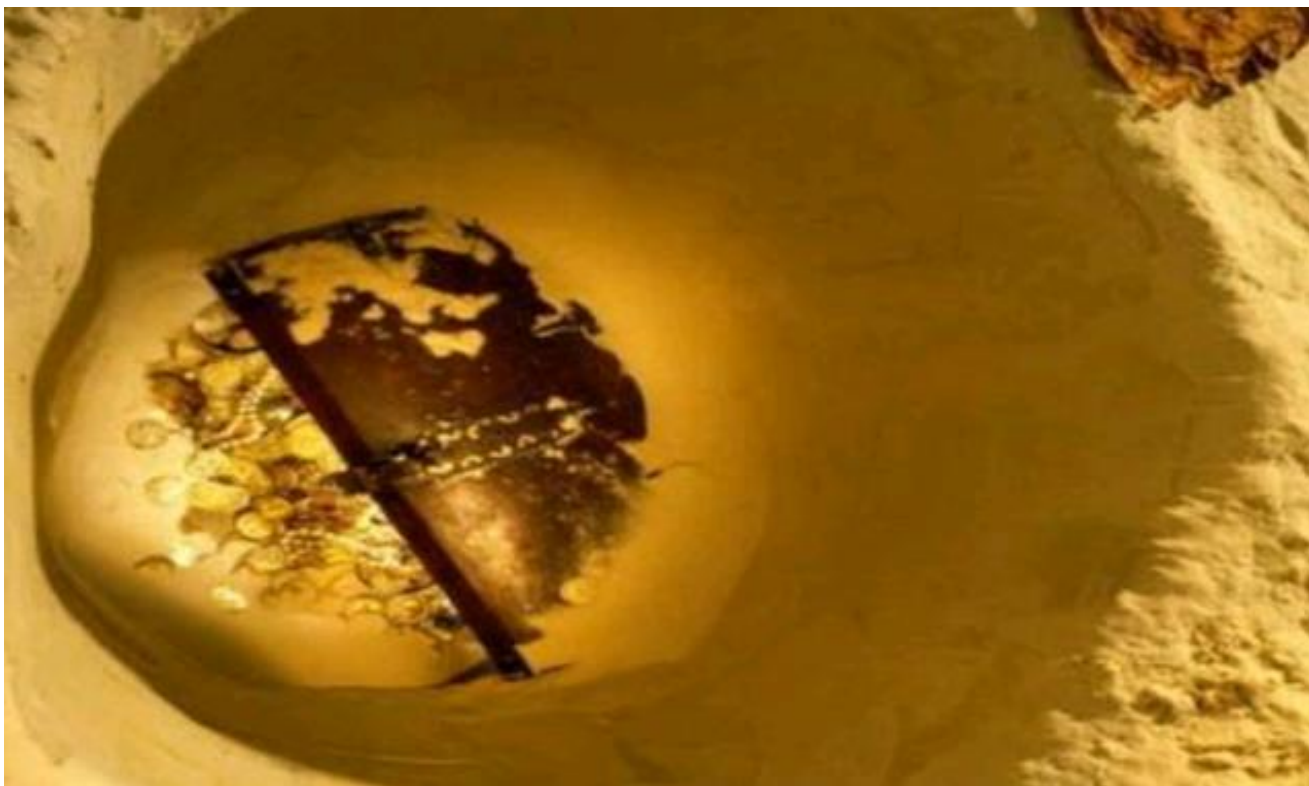


- 实时数据流处理的要求，是区别大数据和传统数据仓库技术，BI技术的关键差别之一；
- 1s 是临界点，对于大数据应用而言，必须要在1秒钟内形成答案，否则处理结果就是过时和无效的；

# 价值密度低

---

挖掘大数据的价值类似沙里淘金，从海量数据中挖掘稀疏但珍贵的信息，就是数据分析与数据挖掘。





# 价值密度低

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



# 大数据发展的历程

“大数据”一词在1980年[美]著名未来学家阿尔文·托夫勒著的《第三次浪潮》书中将“大数据”称为“第三次浪潮的华彩乐章”。

2008年9月《自然》杂志在推出了名为“大数据”的封面专栏。从互联网技术、生物医学等方面，探讨了大数据的研究。

2010年2月，肯尼斯·库克尔在《经济学人》上发表了长达14页的大数据专论《数据，无所不在的数据》。

2011年2月，《科学》杂志推出专刊《处理数据》，讨论了科学研究中的大数据对科学研究的重要性。

“大数据时代已经到来”出现在2011年6月麦肯锡发布了关于“大数据”的报告。大数据的概念，后逐渐受到了各行各业关注。



# 国外发展状况

---

2012年3月29日，美国发布《大数据研究与发展计划》，将大数据的研究和发展上升为国家战略层次。之后，12个联邦部门启动开展了82个大数据相关项目，涵盖了国防、国土安全、国家安全、能源、医疗卫生、食品药物、航空航天、人文社会科学、地质勘查等众多领域，美国希望借助大数据技术实现这些领域的技术突破。

2013年2月法国政府发布《数字化路线图》，列出5项将会大力支持的战略性高新技术，其中一项就是大数据。

2013年10月31日，英国发布《把握数据带来的机遇：英国数据能力战略》，战略旨在促进英国在数据挖掘和价值萃取中的世界领先地位。

2013年6月，日本公布了新的IT战略——《创建最尖端IT国家宣言》，全面阐述了2013~2020年期间以发展开放公共数据和大数据为核心的日本新IT国家战略。



# 国内发展状况

---

## 政府层面

2011年12月，工信部发布的物联网十二五规划上，把信息处理技术作为4项关键技术创新工程之一被提出来，其中包括了海量数据存储、数据挖掘、图像视频智能分析，这些是大数据的重要组成部分。

2013年称为“大数据元年”，大数据已深刻地影响着人类社会的各个领域，带来了信息领域在新时代的革命。

2014年，“大数据”首次出现在当年的《政府工作报告》中。“大数据”旋即成为国内热议词汇。

2015年，国务院正式印发《促进大数据发展行动纲要》，标志着大数据正式上升至国家战略层面。

2016年，国家大数据战略作为“十三五”十四大战略之一，首次被写进五年规划中，大数据创新应用向纵深发展。

2017年，《大数据产业发展规划（2016-2020年）》正式发布，全面部署“十三五”时期大数据产业发展工作，推动大数据产业健康快速发展。

# 浙江省发展状况

---

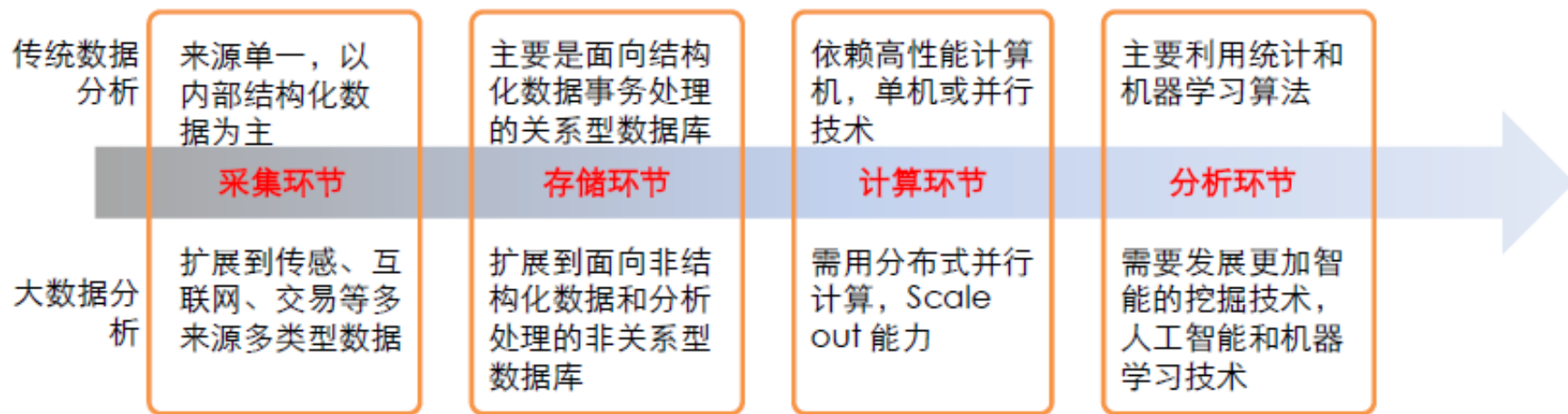
2015年10月14日在2015杭州·云栖大会上时任浙江省省长李强，分享了“数据充满机遇，云端决定未来”的观点，并首次提出“云上浙江”“数据强省”的新概念。

浙江在2016年2月出台《浙江省促进大数据发展实施计划》，提出“建设数据强省，助力经济社会转型升级，推动政府治理和公共服务能力现代化”，成为全国首个出台大数据产业发展计划的省份。“数据强省”成为浙江发展的新目标。

# 大数据处理技术

## 大数据对传统数据处理技术体系提出挑战

大数据具备数据量大、数据类型多、数据处理速度要求高和价值密度低的特点，传统分析系统架构（RDBMS（关系型数据库管理系统）+ 小型机 + 高端阵列模式）下，传统数据库无法支撑海量数据（如100TB以上，性能下降）、非结构化数据，现有的架构无法线性扩展且成本高昂。



# 大数据关键技术

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 <b>机器学习</b> 和 <b>数据挖掘算法</b> ，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

# 大数据计算模式及处理工具

大数据技术是指从各种类型的数据中快速获取有价值信息的技术。

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

主流的三大分布式计算系统：Hadoop，Spark和Storm。

# 大数据的应用领域

---

大数据技术已经应用于各个行业，包括金融、汽车、餐饮、电信、能源和娱乐等。

**制造业：**利用工业大数据提升制造业水平，包括产品故障诊断与预测、分析工艺流程、改进生产工艺，优化生产过程能耗、工业供应链分析与优化、生产计划与排程。

**金融行业：**大数据在高频交易、社交情绪分析和信贷风险分析三大金融创新领域发挥重大作用。

**汽车行业：**利用大数据和物联网技术的无人驾驶汽车，在不远的未来将走入我们的日常生活。

**互联网行业：**借助于大数据技术，可以分析客户行为，进行商品推荐和针对性广告投放。

**餐饮行业：**利用大数据实现餐饮O2O模式，彻底改变传统餐饮经营方式。

**电信行业：**利用大数据技术实现客户离网分析，及时掌握客户离网倾向，出台客户挽留措施。

# 大数据的应用领域

---

**能源行业：**随着智能电网的发展，电力公司可以掌握海量的用户用电信息，利用大数据技术分析用户用电模式，可以改进电网运行，合理设计电力需求响应系统，确保电网运行安全。

**物流行业：**利用大数据优化物流网络，提高物流效率，降低物流成本。

**城市管理：**可以利用大数据实现智能交通、环保监测、城市规划和智能安防。

**生物医学：**大数据可以帮助我们实现流行病预测、智慧医疗、健康管理，同时还可以帮助我们解读DNA,了解更多的生命奥秘。

**体育娱乐：**大数据可以帮助球队训练，决定投拍哪种题材的影视作品，以及预测比赛结果。

**安全领域：**政府可以利用大数据技术构建起强大的国家安全保障体系，企业可以利用大数据抵御网络攻击，警察可以借助大数据来预防犯罪。

**个人生活：**大数据还可以应用于个人生活，利用与每个人相关联的“个人大数据”，分析个人生活行为习惯，为其提供更加周到的个性化服务。

# 云计算的定义

---

**美国国家标准和技术研究院：**云计算是一种能够通过网络以便利的、按需付费的方式获取计算资源，这些资源来自一个共享的、可配置的资源池，并能够以最省力和无人干预的方式获取和释放。



**伯克利云计算白皮书：**云是一个包含大量虚拟资源的资源池，包括硬件、开发平台和I/O服务，这些资源可根据不同的负载动态地进行配置，资源池通常按照服务等级协议SLA，采用即用即付的模式进行管理。



# 云计算服务

---



# 云计算架构



# 云计算关键技术

---

虚拟化技术是指将一台计算机虚拟为多台逻辑计算机，在一台计算机上同时运行多个逻辑计算机，每个逻辑计算机可以运行不同的操作系统，并且应用程序都可以在相互独立的空间内运行而互不影响。

优点：

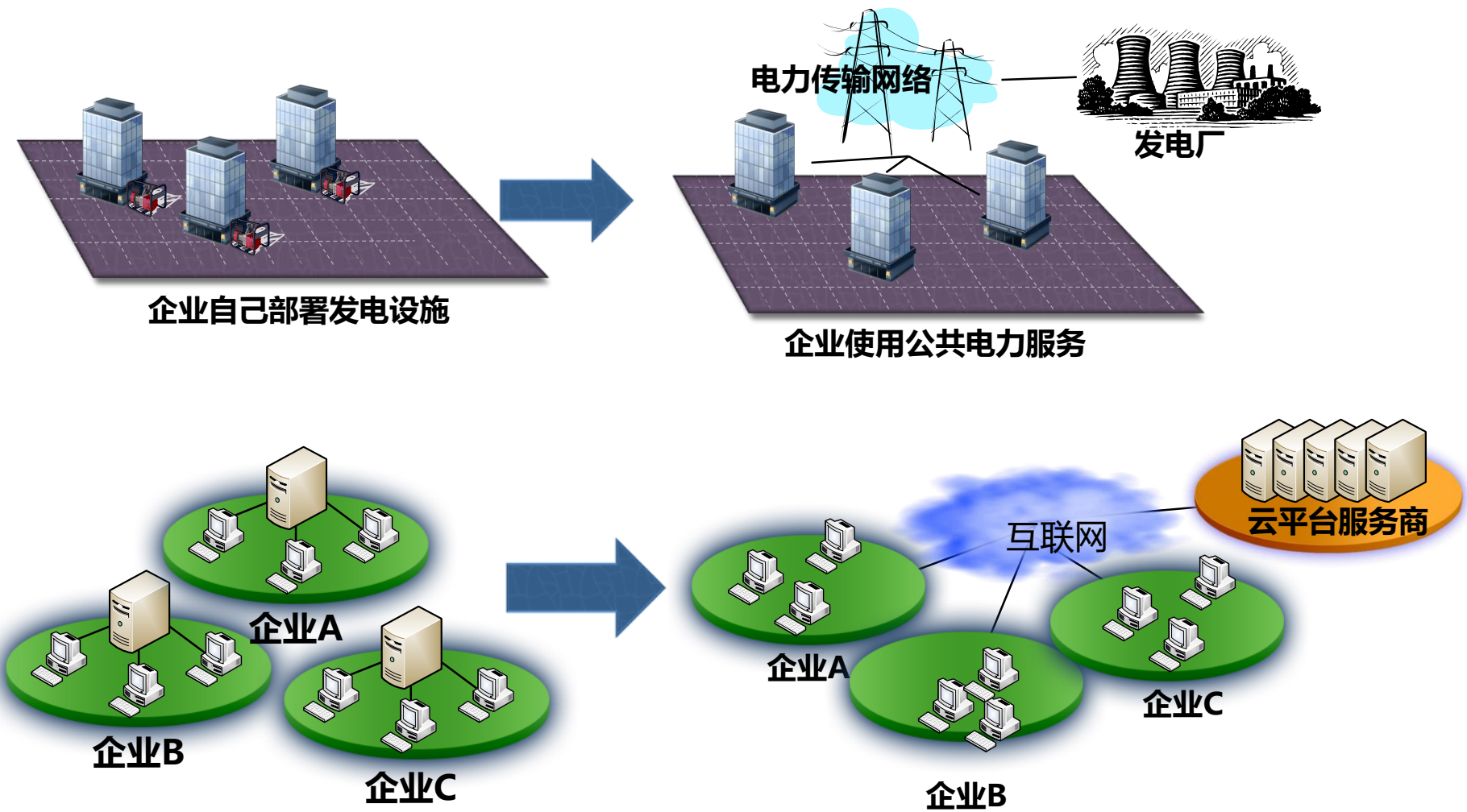
- **提高资源的利用率。**多个操作系统可以同时存在和运行于同一个物理平台上（在单个服务器上有可能同时运行数百个虚拟机器）。
- **有效隔离操作系统和资源。**虚拟机中的操作系统崩溃后恢复比较影容易，并不会对同一个物理平台上的其它操作系统造成响，而且比较容易实现操作系统的数据重放和回滚。

# 云计算核心服务



# 云计算目标

像用电、水一样使用IT

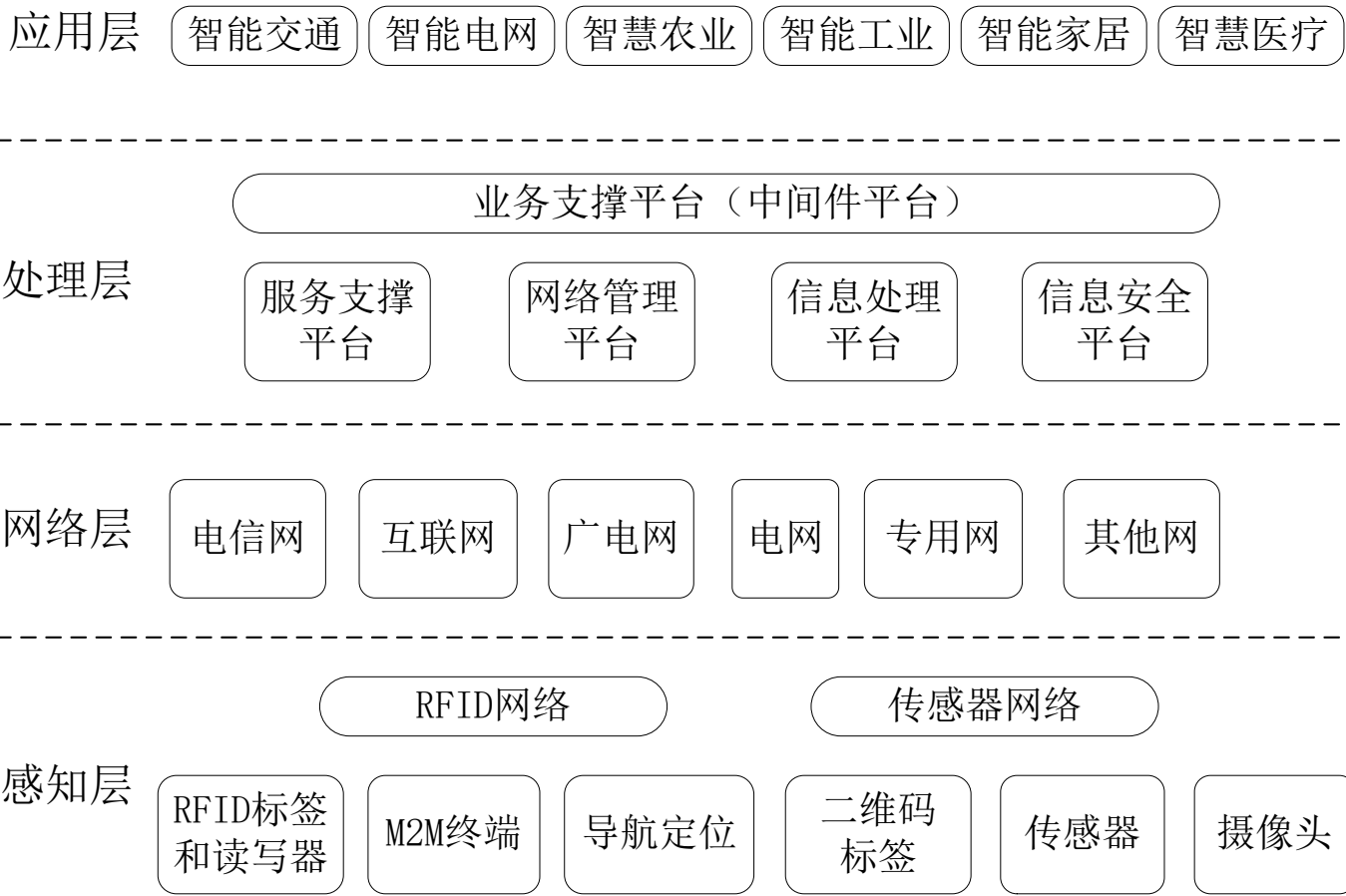


# 物联网的定义

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。



# 物联网的体系架构



# 物联网核心技术——感知层

- 感知识别层位于物联网四层模型的最底端，是所有上层结构的基础。
- 信息生成方式多样化是物联网的重要特征之一。
- 通过感知识别技术，让物品“开口说话、发布信息”是融合物理世界和信息世界的重要一环，是物联网区别于其他网络的最独特的部分。
- 物联网的“触手”是位于感知识别层的大量信息生成设备，既包括采用自动生成方式的射频识别技术RFID、传感器、定位系统等，也包括采用人工生成方式的各种智能设备，例如智能手机、PDA、多媒体播放器、上网本、笔记本电脑等等。

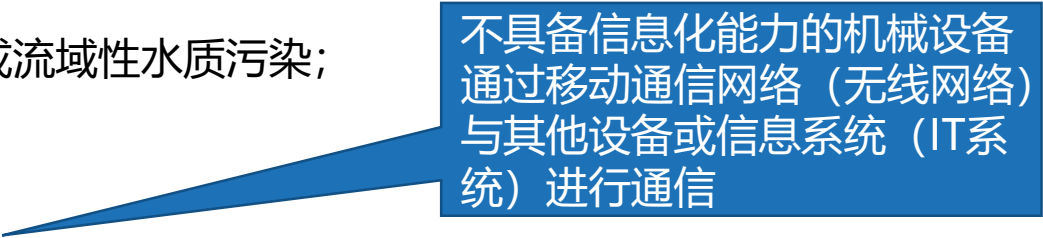




# 物联网的应用方向

---

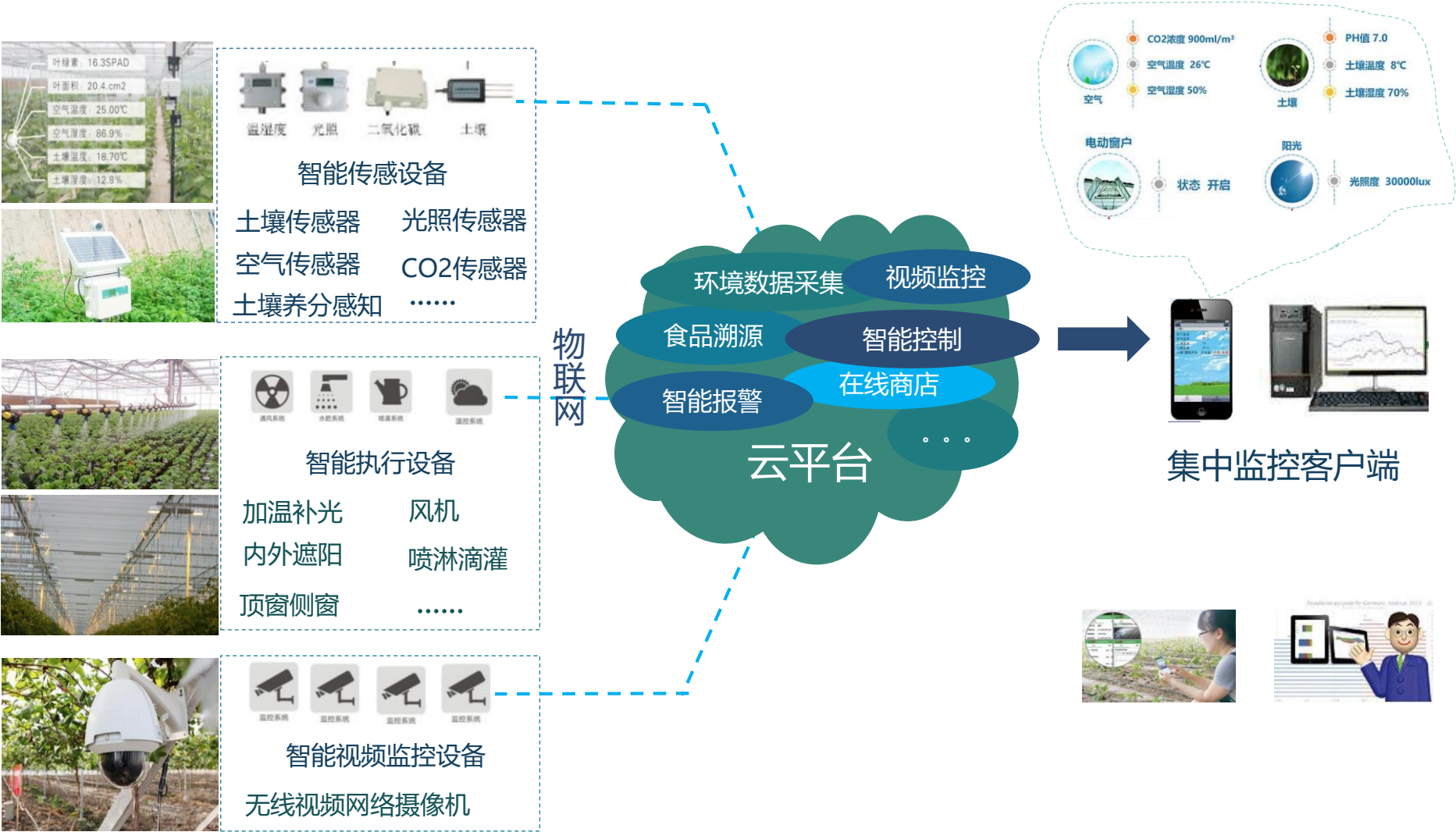
- (1)智能城市，用于城市的数字化管理和安全监控；
- (2)智能交通，包括公交视频监控、智能公交站台、电子票务、车管专家和公交手机一卡通、红绿灯自动控制和交通违章监管等业务；
- (3)智能物流，打造集信息展现、电子商务、物流配载、仓储管理、金融质押、园区安保、海关保税等功能为一体的物流园区综合信息服务平台；
- (4)智能环保，实施对水质的实时自动监控，预防重大或流域性水质污染；
- (5)智能家居，用于各种家庭设备的控制；
- (6) M2M (machine to machine) 应用；
- (7)精准农业，通过实时采集温度、湿度、光照、CO<sub>2</sub>浓度以及土壤温度、叶面湿度等参数，实现对指定设备自动关启的远程控制；
- (8)智能医疗，用于病人监测护理等服务。



不具备信息化能力的机械设备  
通过移动通信网络（无线网络）  
与其他设备或信息系统（IT系统）  
进行通信

# 智慧农业大棚

通过智能硬件、物联网、大数据等技术，采集环境和植物生长数据，为智能控制和创造生长环境提供条件，实现“科学指导生态轮作和智能化管理”。



# 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。

