大数据分析与智能计算

Fall 2023

Instructor Dr. Yu Tang, Dr. Di Lin
Lab Advisor Ms. Yang and Ms. Li
Office 401 SISE Building, Sand River Campus
Phone (028) 83208057
Faculty Email yutang@uestc.edu.cn

上课时段	周二	周三
10:20 – 11:55	二教110	二教110
19:30 – 21:05	二教107	二教107

课程目标

课程定位：一门培养软件工程专业学生大数据分析与智能计算能力的专业核心课程

课程目标：

CO1：建立大数据和云计算的基本概念，了解大数据分析和智能计算的原理，培养学生在民族复兴和创新发展中国新技术方面的责任感

CO2：理解分析算法的特点，并掌握基于Python和Java语言的数据分析方法

CO3：理解大数据计算架构及云计算系统的设计原理及实现方法

CO4：初步具备按照业务需求进行数据分析算法设计和大数据计算系统开发实现的能力

毕业要求指标点

本课程支撑的毕业要求指标点

GR3.3 能够针对复杂软件工程问题，设计满足特定需求的总体设计和详细设计，体现创新意识

GR3.4 能够集成单元过程进行软件系统流程设计，对流程设计方案进行优选

GR6.1 掌握至少一个应用领域中软件工程技术的应用方法和工程实践

课程内容

- 大数据概念
- 大数据计算体系
- 数据采集与建模
- 大数据分析算法
- 大数据处理技术
- 数据可视化
- Hadoop计算体系
- HDFS/HBase存储架构
- MapReduce计算模型
- 图并行计算框架
- 流计算
- 内存计算
- 云计算概念
- 云计算架构
- 云系统开发技术
- 开源云计算平台

课程实验

- 实验1 Hadoop单机安装并使用MapReduce实现Wordcount实例
- 实验2 Spark、PySpark、Python、Jupyter（Pycharm）安装部署
- 实验3 针对DataExpo2009数据集，进行航空公司航班延误分析预测的大数据计算及算法实现，竞争性指标结果

教学模式

- 讲座 + 课外阅读 + Projects
- 需具有OO编程，O/S，数据库基础
- 强调动手能力，上机操作并演示
- 大量课外阅读：在线技术性文档
- 开放性教学：课堂讨论+课外交流

评分标准

Project	40%
期末考试	60%
Total	100%

85~100	**A**
75~84	**B**
60~74	**C**
< 60	**Fail**

教材与参考书

教材

汤羽、林迪等编著，**《大数据分析与计算》**，清华大学出版社，2021年10月第7次印刷

刘鹏主编，**《云计算》（第三版）**，电子工业出版社，第3版，2015

教材与参考书

参考书

麦金尼 编著，**《利用Python进行数据分析》** 机械工业出版社， 第1版，2014

刘凡平 编著，**《大数据时代的算法: 机器学习、人工智能及其典型实例》** 电子工业出版社，第1版，2017

（美）Rachel Schutt, Cathy O’Neil 著，**《数据科学实战》**，人民邮电版社，2015年3月第1版

高彦杰，**《 Spark大数据处理：技术、应用与性能优化》**，机械工业出版社，2014年11月

Lecture 1 大数据计算概论

1.1 大数据概念

1.2 大数据技术特征

1.3 云计算概念

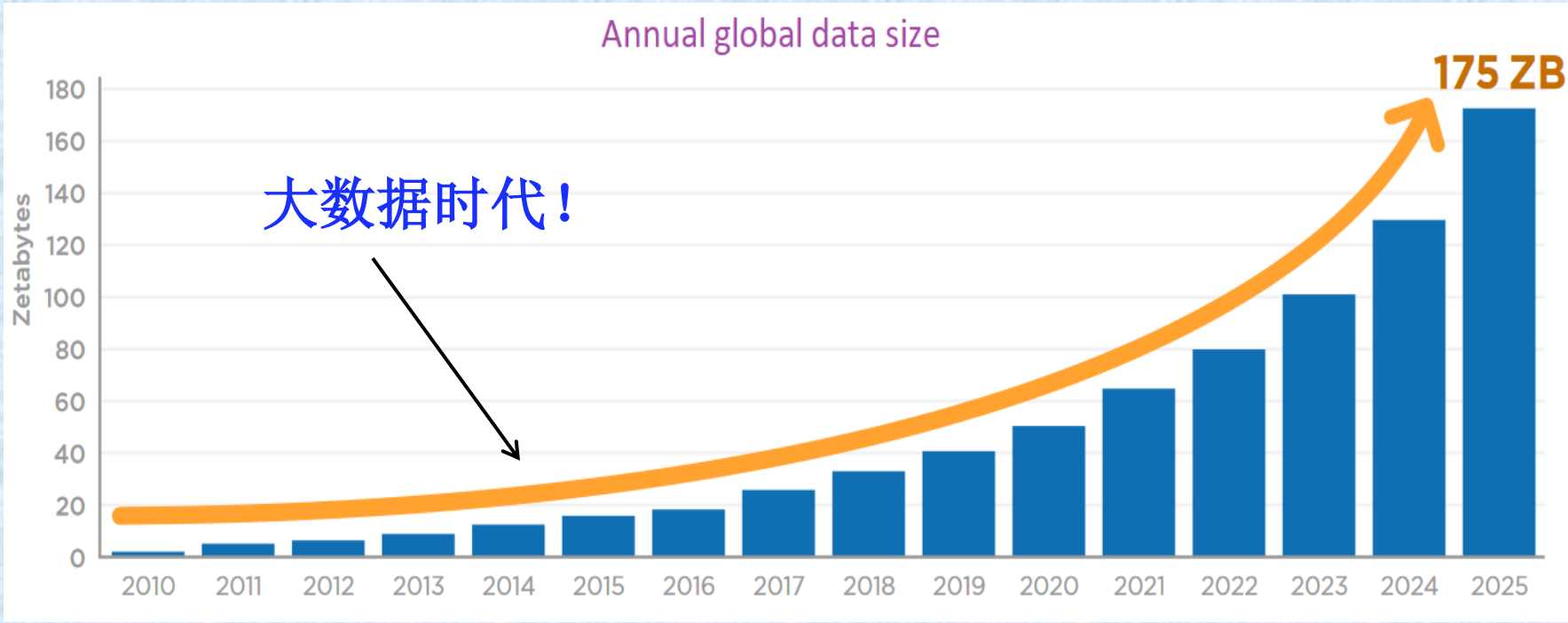
1.1 大数据概念

- 数据是什么？
- 数据科学是什么？
- 大数据基本属性是什么？

全球视野下的大数据：机遇与挑战

“黄河之水天上来”！

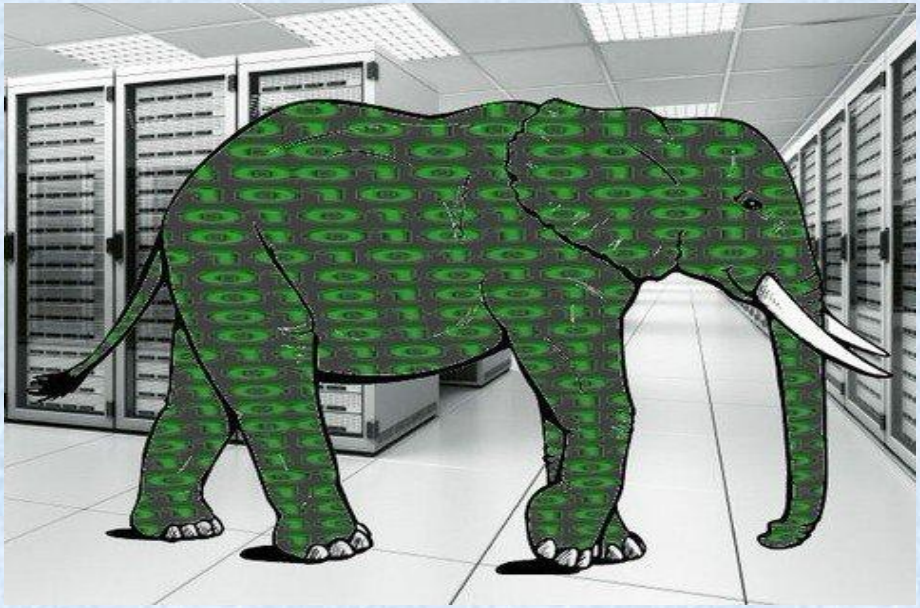
- Facebook每天处理80亿条信息
- Google每天完成10亿次查询
- 全世界的信息量以每两年翻番的速度增长
- 2011年全球数据量为1.8ZB，IDC预测2015年达到8ZB, 2020年更将达到35ZB！



$1\ ZB = 10^3\ PB = 10^6\ TB = 10^9\ GB$

■ 什么是大数据（Big Data）？

- Volume: 数据量异常庞大，一般达到PB量级
- Variety: 数据呈异构化，数据来源呈多样性
- Velocity: 数据处理要求时效性
- Value: 单个数据无价值，但大规模数据拥有巨大价值



■ 什么是大数据？（续）

- 数据种类的多样性：文字、语音、图片、视频、信息等
- 数据对象的多样性：个人信息、个人数据、商业服务数据、社会公共数据、自然界数据、物质世界的数据
- 数据来源的多样性：在数据层面打破现实世界的界限，多家公司的共享替代一家公司的数据

■ 大数据已上升到21世纪国家战略的高度

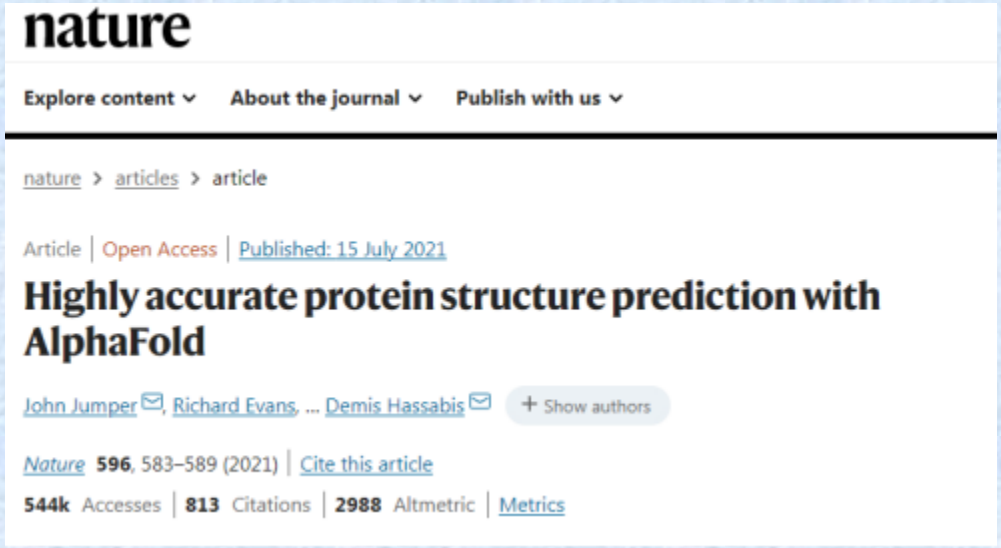
2012年3月美国奥巴马政府宣布推出“*大数据的研究和发展计划*”，包括

- 美国国家科学基金（NSF）
- 美国国家卫生研究院（NIH）
- 美国能源部、美国国防部
- 美国国防部高级研究计划局、美国地质勘探局等6个联邦政府部门

■ 大数据案例：Google AlphaFold

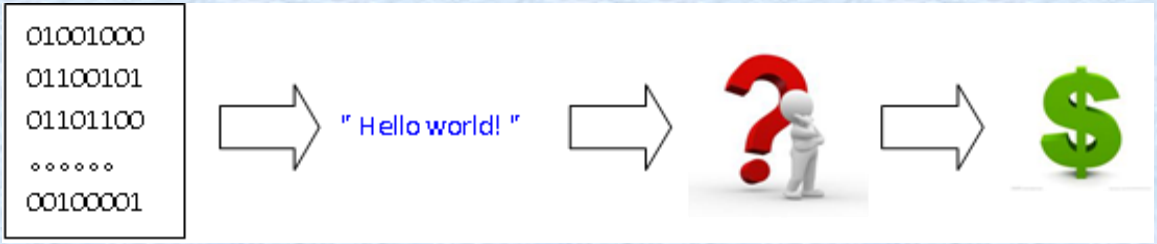
2020年，在第14届国际蛋白质结构预测竞赛（Critical Assessment of Protein Structure Prediction，CASP）上，AlphaFold2成功根据基因序列预测了生命基本分子——蛋白质的三维结构，取得了中位分数为92.4（满分100分），比第二名高25分，打败了所有竞争对手。

2021年Deep Mind宣布：“AlphaFold2成功解开了一个困扰人类长达50年之久的生物学难题——蛋白质折叠问题。”2021年7月15日，**AlphaFold2的论文在Nature上发表，并在Github上将AlphaFold2的代码开源，以及上线可搜索的物种蛋白质组数据库**。中国科学院院士施一公认为：AlphaFold2是人工智能对科学领域最大的一次贡献，也是人类在21世纪取得的最重要的科学突破之一。



大数据概念——数据的定义

- 数据的定义
  - 数据的基本定义
  - 计算机学科中数据的定义
- 数据的多样化
  - 数据的形式多样化
  - 数据的来源多样化
  - 数据的范围多样化
- 数据转换过程
  - 数据-信息-知识-价值转换模型



大数据概念——基本属性

- Volume：大数据的超大规模
  - 规模体现
  - 带来的影响：
    - 数据存储架构：
      - 基于行-键表格存储格式的关系型数据库？
      - 基于分布式文件系统的分布式数据库！
    - 计算模型：
      - 离线批处理计算框架（ MapReduce）
      - BSP图并行计算框架（Pregel、Hama）
      - 交互式计算模型
      - 大内存计算系统

大数据概念——基本属性

- **Variety：大数据来源多样性与异构性**
  - 大数据类型划分：
    - 依结构特征划分
    - 依时效性划分
    - 依关联特性划分
    - 依数据类型划分
    - 依数据来源划分
  - 带来影响：
    - 数据存储、管理和快速查询异常困难

大数据概念——基本属性

- Value：价值低密度特性
  - 区别于传统数学统计学方法的关键之处

	传统数学统计学	大数据分析计算方法
处理对象	局部数据或数据子集	以数据整体或完整数据集作为处理对象
处理方法	基于抽样调查的随机分析方法	机器学习方法 通过数据的积累来训练和改进算法和计算程序
结果正确性	取决于随机抽样模型产生的数据集的代表性	处理数据量越大，计算结果越越优化

1.2 大数据技术特征

- 大数据算法特性
- 大数据计算系统特性
- 大数据开发技术特性

大数据算法特性

	大数据计算	传统统计学	优势
样本空间	整个数据集	基于独立同分布原理抽取样本集	避免样本失真
计算方法	机器学习方法	按照固定数学模型进行预测	预测结果的精度改进是一个动态过程

大数据计算系统特性

	大数据计算系统	传统数据库系统	优势
基础模型	分布式文件系统 NoSQL非关系型数据库	关系型模型	支持非结构化或异构数据的存储和处理 支持分布式系统部署 支持超大规模数据集完成快速查询操作
存储格式	基于键值对的列存储格式	基于主键的行存储格式	更优的查询效率 更好的对计算模型的支持

大数据计算系统特性

某大学学生总数
N=30000

数据库中每个学生相关值域数量m=50

从数据库中搜出并计算某一专业学生（含不同年级）某一门课的平均成绩？

关系型数据库：

从数据库总表中搜出满足上述条件的学生记录，操作次数是O(N)量级

对搜出的每一条学生记录完成该门课程成绩的读取，操作次数是O(m)

总操作次数为O(N) * O(m) 量级，最坏情况下需要操作 30000 x 50 = 1500000次！

某大学学生总数
N=30000

数据库中每个学生相关值域数量m=50

从数据库中搜出并计算某一专业学生（含不同年级）某一门课的平均成绩？

NoSQL数据库：

所有学生的成绩都存入树状结构的某一分枝

搜索进入该门课的分枝
（最坏情况下查询次数2000）

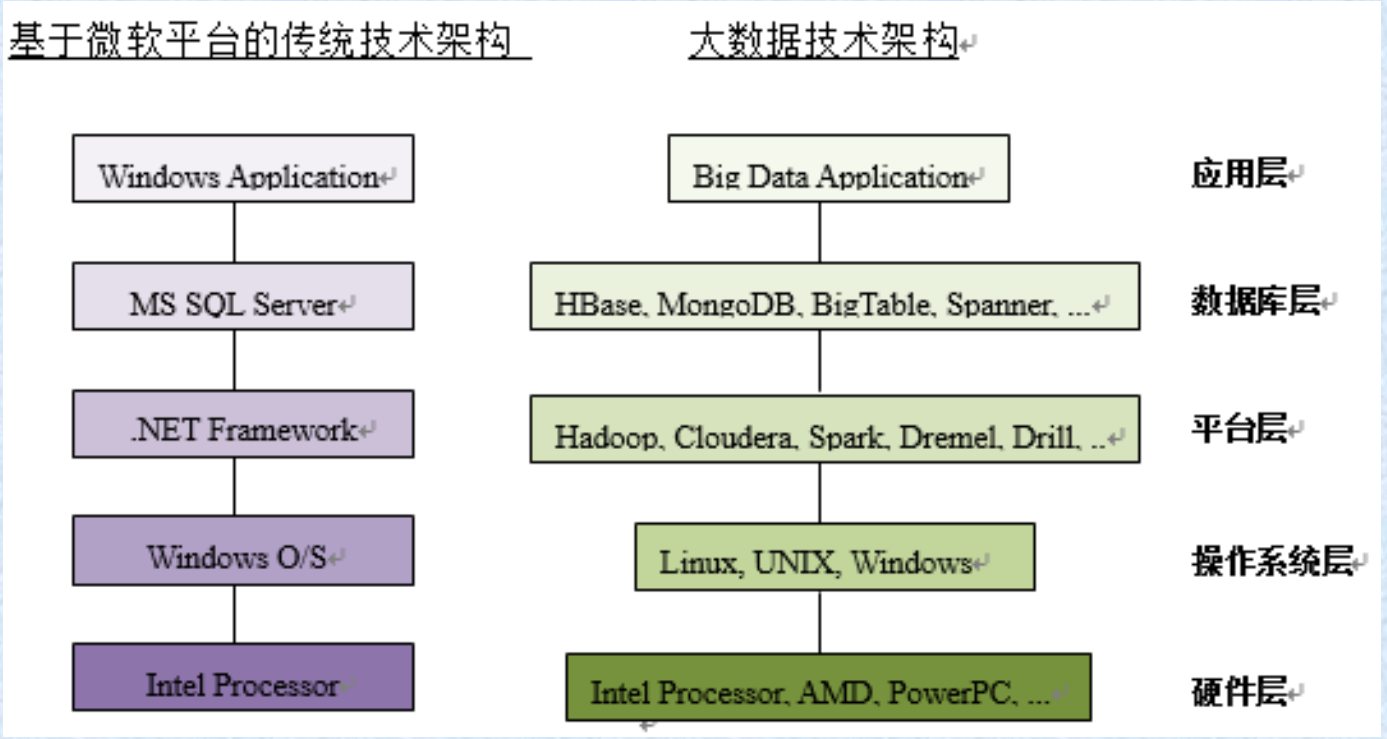
在该分枝内搜索该专业
（最多查询次数100）

完成符合条件的学生成绩的读取（最多读取1000次）

总的操作次数为：
2000 + 100 + 1000 = 3100次

大数据开发技术特性

大数据计算系统	传统数据库系统	优势
多层次的分层结构	基于某一平台和某一标准的线性结构	在同一平台上尽可能多的兼容或集成不同的软件开发工具



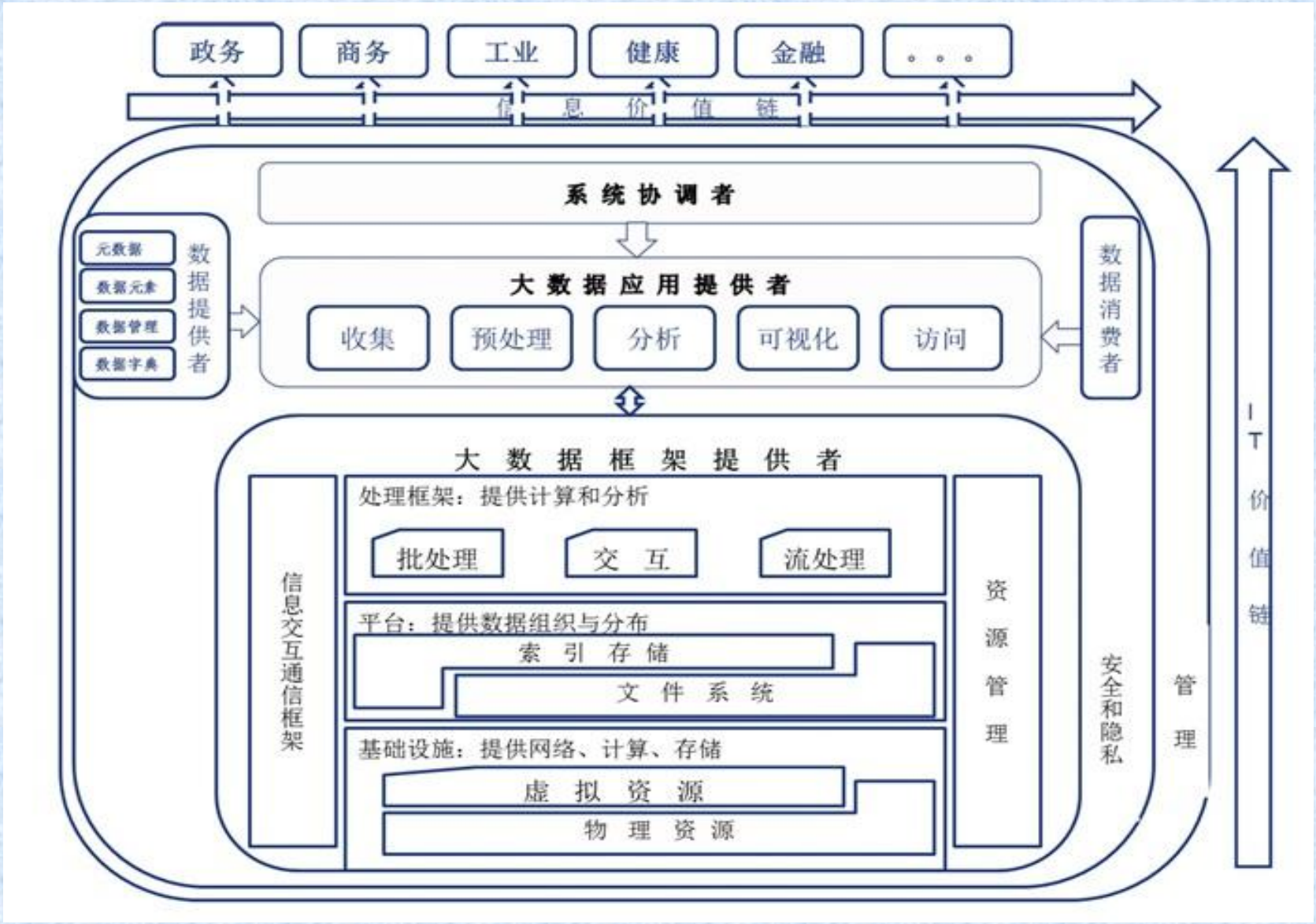
大数据计算技术标准

大数据计算技术标准

- 大数据技术架构参考模型
- 大数据计算体系主要角色
- 大数据标准体系框架

大数据计算模式

- 主要计算模式
- 各计算模式特性与优劣
- 大规模并行处理模式



大数据技术架构参考模型

大数据计算技术标准

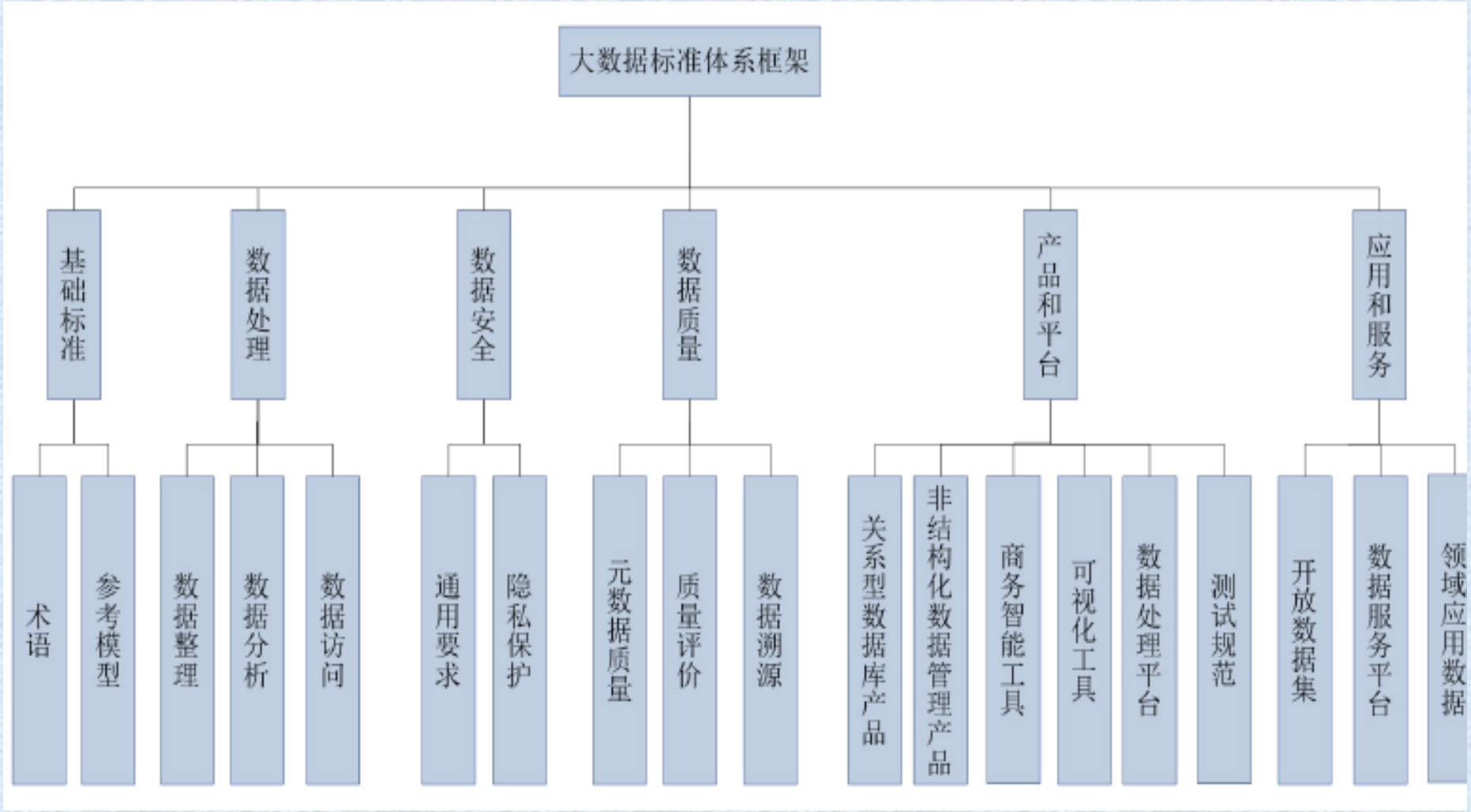
- 大数据技术架构参考模型基于两个维度组成：信息链（垂直方向）和价值链（水平方向）
- 信息链维度：通过数据采集、集成、分析、使用结果来实现价值
- 价值链维度：通过为大数据应用的实施提供拥有或运行大数据的网络、基础设施、平台、应用工具以及其他IT 服务来实现价值

大数据计算技术标准

大数据计算体系主要角色

- 系统领导者
- 数据提供者
- 安全和隐私角色
- 大数据应用提供者
- 大数据基础框架提供者
- 数据消费者
- 管理角色
- 安全及隐私管理角色

大数据标准体系


大数据标准体系框架
基础标准
术语
参考模型
数据处理
数据整理
数据分析
数据访问
数据安全
通用要求
隐私保护
数据质量
元数据质量
质量评价
数据溯源
产品和平台
关系型数据库产品
非结构化数据管理产品
商务智能工具
可视化工具
数据处理平台
测试规范
应用和服务
开放数据集
数据服务平台
领域应用数据

大数据计算技术标准

大数据标准体系框架组成

- 基础标准
- 数据处理标准
- 数据安全标准
- 数据质量标准
- 产品和平台标准
- 应用和服务标准

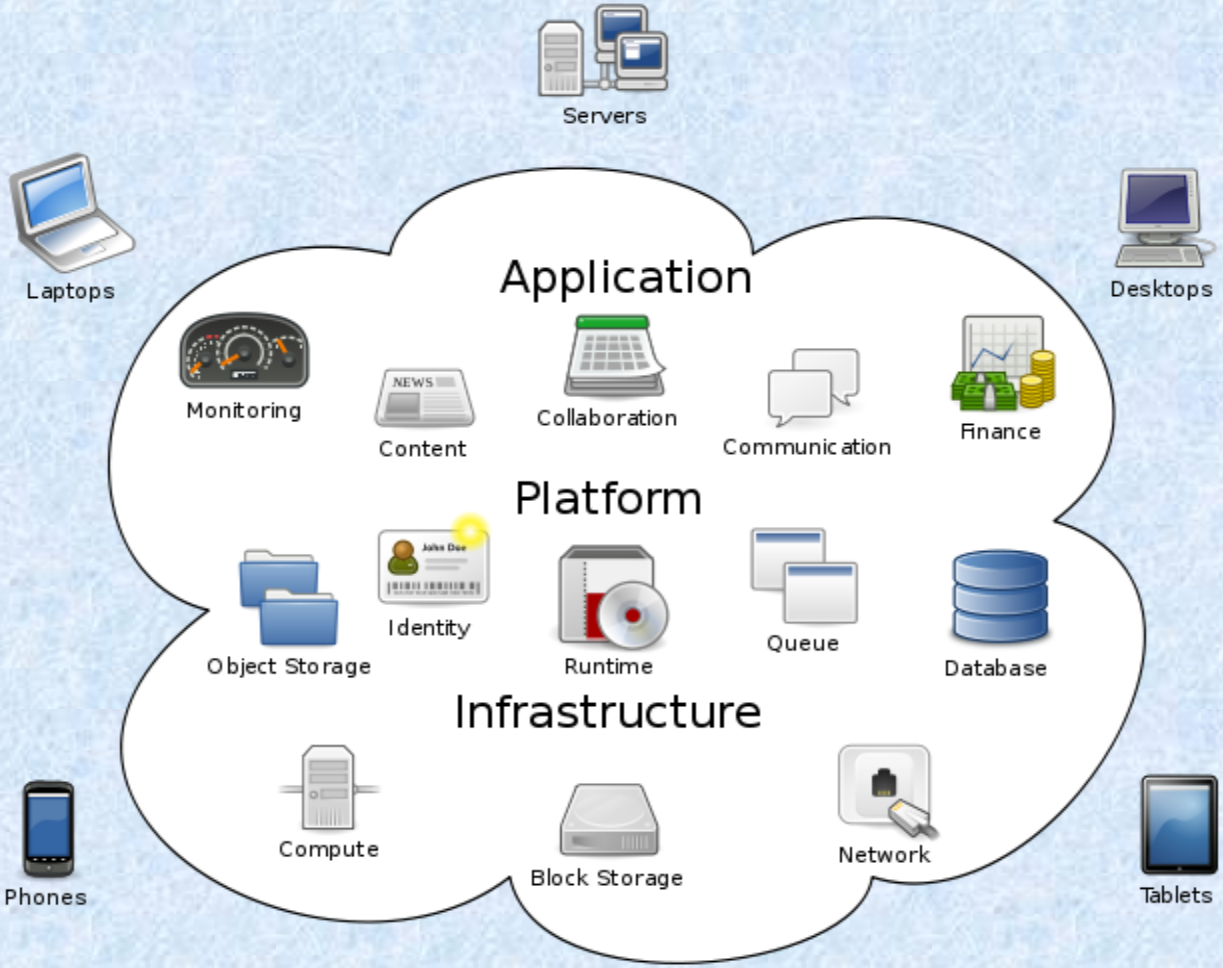
1.3 云计算概念

What is Cloud Computing?

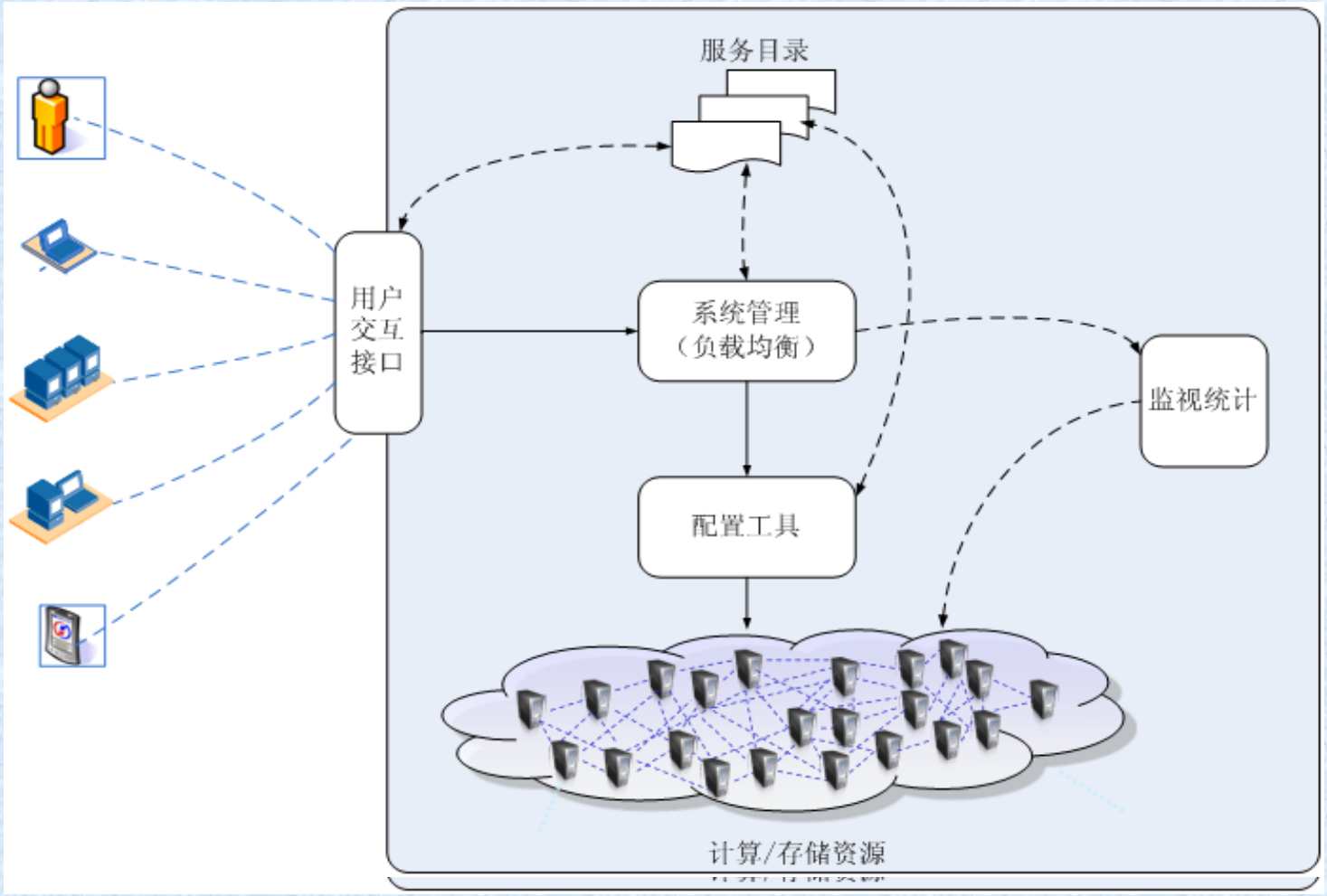
Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction --- *NIST*

云计算是一种商业计算模型。它将计算任务分布在大量计算机构成的资源池上，使各种应用系统能够根据需要获取计算力、存储空间和各种软件服务 --- *刘鹏*

云计算系统架构


Servers
Laptops
Desktops
Application
Monitoring
Content
Collaboration
Communication
Finance
Platform
Identity
Object Storage
Runtime
Queue
Database
Infrastructure
Compute
Block Storage
Network
Phones
Tablets
Cloud Computing

云计算部署架构



云计算发展历史

需求推动、技术进步、商业模式变革

Google

超过4万台服务器

约需要20P（2千万G）以上空间

YouTube 超过10亿个视频短片，每秒有1万人次观看（每天超过10亿人次）

视频

facebook 保存有800亿张照片，每秒有60万人次访问

音乐

照片

flickr

文本

twitter

智慧地球

云计算：从现在开始

RFID、传感器

物联网

万亿级

上网本、智能手机

移动互联网

40亿

互联网：90年代初

20亿

笔记本电脑

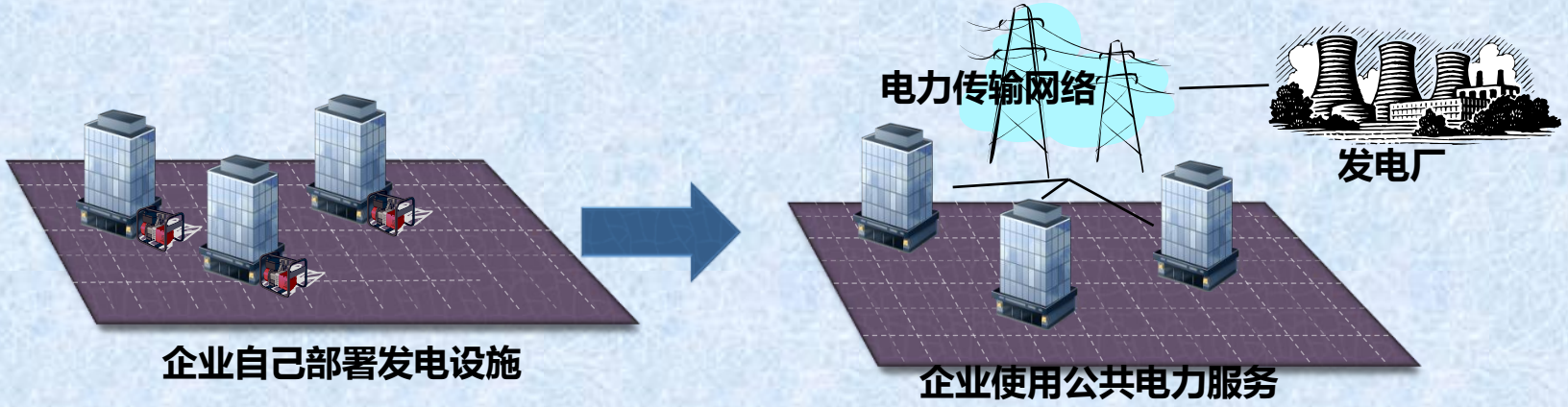
桌面电脑

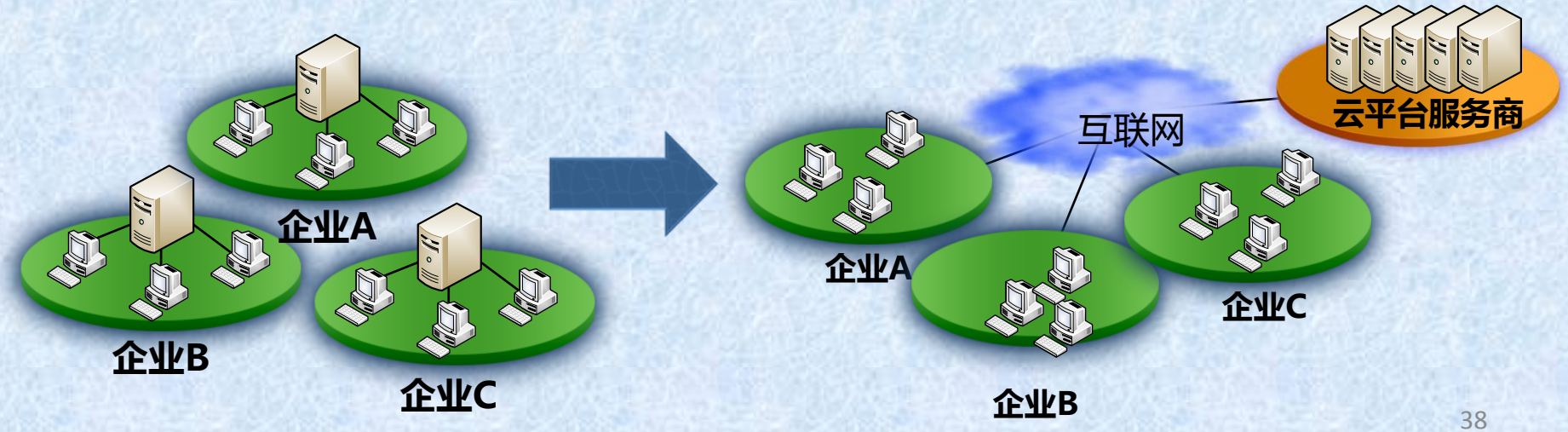
个人电脑：80年中期

10亿

云计算目标

像用电、水一样使用IT资源





云计算的组成

云计算的组成可以分为六个部分，它们由下至上分别是：

- 基础设施（Infrastructure）
- 存储（Storage）
- 平台（Platform）
- 应用（Application）
- 服务（Services）
- 客户端（Clients）


Why Cloud?

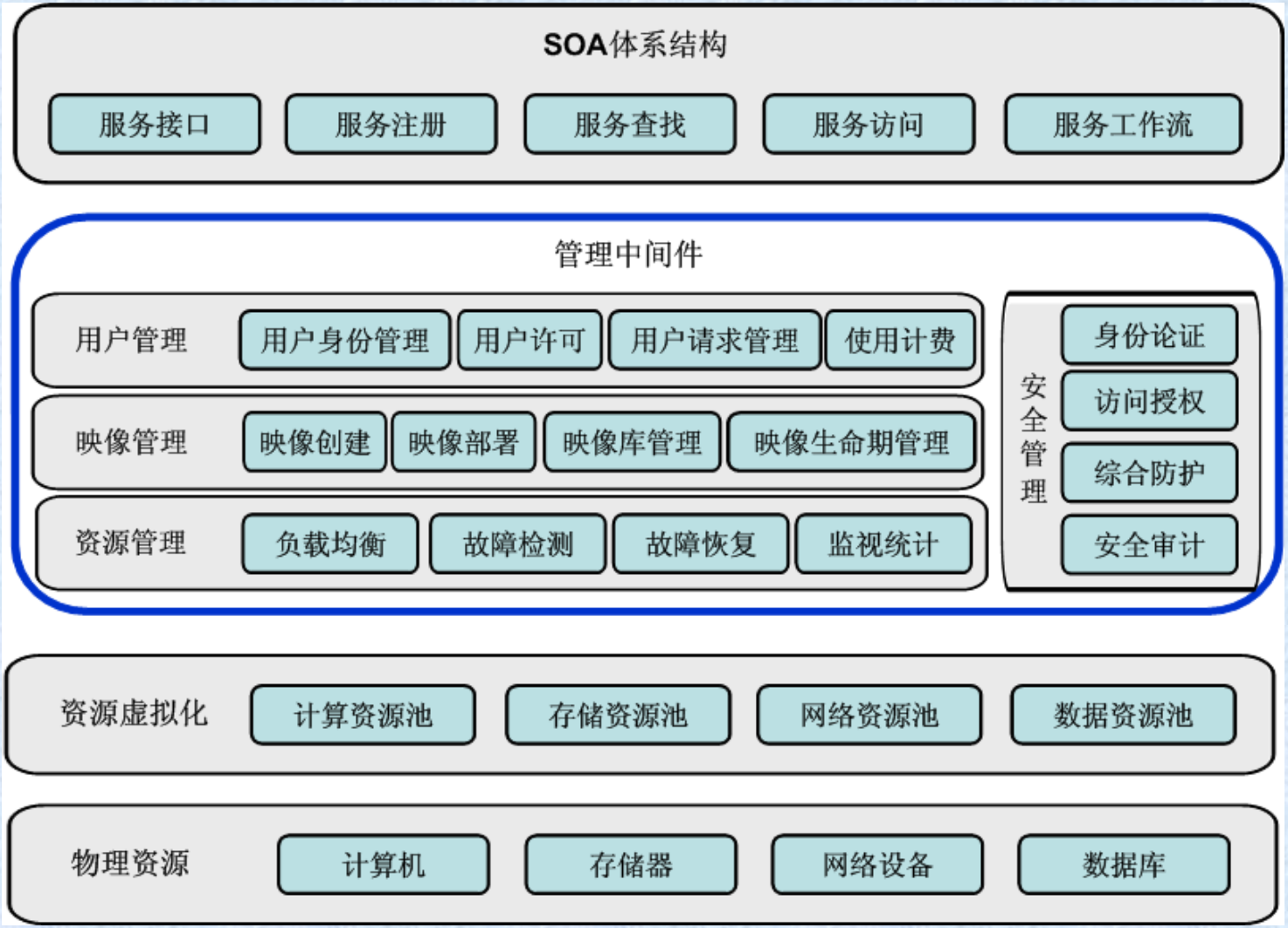
Advantages of Cloud Computing

- high availability
- high scalability – extra large-scale
- on-demand service
- economic
- high usability via visualization

Layered Cloud Software Model

- cloud application layer
- virtualized execution environment
- integrated service/function layer
- computing resource scheduling layer
- data storage/management layer
- IT infrastructure layer (servers, storage, network, etc.)

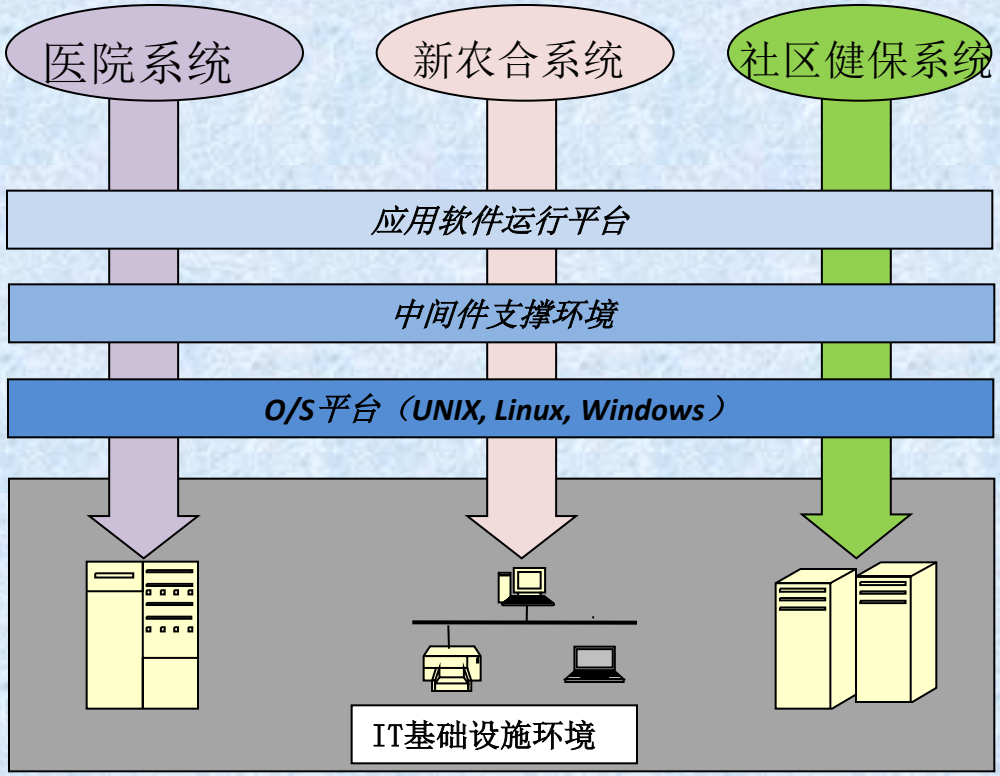
云计算系统架构


SOA体系结构
服务接口
服务注册
服务查找
服务访问
服务工作流
管理中间件
用户管理
用户身份管理
用户许可
用户请求管理
使用计费
映像管理
映像创建
映像部署
映像库管理
映像生命期管理
资源管理
负载均衡
故障检测
故障恢复
监视统计
安全管理
身份论证
访问授权
综合防护
安全审计
资源虚拟化
计算资源池
存储资源池
网络资源池
数据资源池
物理资源
计算机
存储器
网络设备
数据库

云计算应用案例

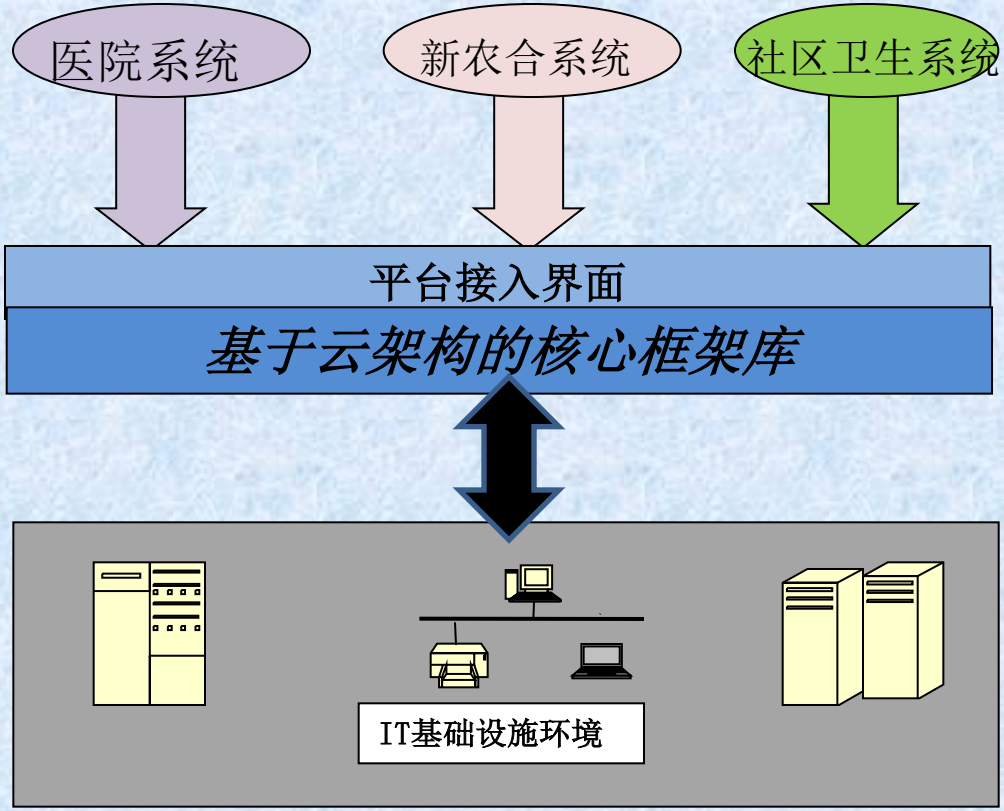
现有系统的分散式、条块分割体系 –“电线杆式”

- 互不联通、数据孤岛
- 条块分割，难于实现资源动态调配
- 应用软件与下层硬件平台绑定，硬件更新影响上层系统，维护困难



基于云架构的区域卫生信息框架库

- 上层业务系统不依赖于下层硬件平台，便于维护、升级
- 计算资源的管理调配由平台统一管理、可实现优化调配
- 软件采用组件式框架库设计，可组装使用，降低开发维护成本



Thanks !