汇总

摩尔定律：半导体；吉尔德：网络发展；麦特卡夫：网络用户和经济价值

四层两域

IaaS EC2, PaaS AWS, DaaS Hadoop, SaaS

SDN

ISA, JVM, .NET

VMM/Hypervisor, 半虚拟化，完全虚拟化

Docker, 共享内核, Kubernets

分布式系统是软件系统，透明性，开放性，可拓展性

分布式计算，MapReduce, Spark, OOM, CORBA, DCOM, Java/RMI, SOM, Web Services, CORBAA, UBM, VJM, Storm

分布式存储，可靠，HBase

**分布式文件系统**（非结构化数据），NFS, Andrew AFS, Kass, GFS (Google File System), Master, Chunkserver, Client, Client-Server, HDFS (Hadoop DFS), NameNode, DataNode

分布式数据库系统，结构化数据，元数据，DBMS，冗余，ACID, CAP, BASE

MEMS

IEEE1451，UART，SPI

1D/二维码，IC卡，RFID，OCR，ASR

**WSN**，**自组织网络**，MAC，S-MAC，T-MAC，能量感知路由，查询路，**ZigBee**，IEEE802.15.4

MQTT, COAP, AMQP, STOMP

WPAN，TG4/ZigBee，TG3/UWB，NFC, 802.15

WLAN，Wi-Fi, 802.11

WMAN，LoRa, 802.16

LPWAN，NB-IoT

包结构，信道编码，数据集成，数据消减，无线传感网，ETL，NewSQL（可扩展），NoSQL（非关系），MapReduce，Spark, 分布式并行处理框架

定义层，实例层

描述性分析，诊断性分析，预测性分析，指令性分析

全样而非抽样，效率而非精确，相关而非因果，以数据为中心，“我为人人，人人为我”

Apriori，DHP，DIC

反馈型网络，前馈型网络，单层神经网络，SVM，ANN，DBN，玻尔兹曼机RBM，CNN，RNN，DBN，GAN

Contents

[第一章 云计算 4](#_Toc210161109)

[第二章：物联网 9](#_Toc210161110)

[第三章：数据基础 13](#_Toc210161111)

[第四章：人工智能 17](#_Toc210161112)

# 第一章 云计算

特征

* 虚拟化（核心），服务化，柔性化，个性化，社会化，智能化
* 不包含设备虚拟化

目标

* 成为基础

技术架构（四层两域）

* 业务域
  + 包含四层用来提供资源和服务（解耦）：基础设施、虚拟池、服务、应用
* 管理域

种类

* 公有云，私有云，混合云（VPN，成本控制，新技术引用，安全扩展）

服务模型

* IaaS：
  + 卖出虚拟电脑，其它软件需要自己安装
  + EC2
* PaaS：
  + 提供开发工具齐全的平台
  + AWS
* DaaS：
  + 通过API或查询接口，提供远程访问所需数据
  + Hadoop
* SaaS：
  + 卖出软件
  + 邮箱，网盘

发展历史：因为IO原因，所以进行如下演变

传统架构 → 虚拟化架构 → 融合架构（分布式+cache应对I/O）

**支撑云计算的网络架构技术** **SDN**

SDN (有全局视野，网络虚拟化)

* Switch: 转发规则由流量表flow table指定
* Controller: 动态生成flow table，因为有全局视野

**虚拟化**

* 定义：解耦物理资源和上层软件。
* 粒度：
  + **指令级**（**ISA**）：最灵活
  + **硬件级**（**CPU/内存/IO**）：性能高，复杂
  + **操作系统级**（**OS 内核**）：性能高
  + **运行时库级**（**API/JVM/.NET**）
  + **应用程序级**（**封装**）：隔离性最强，复杂

**虚拟机**：

* VMM/Hypervisor = 虚拟化管理器（硬件虚拟化）
* 半虚拟化：需修改OS
* 完全虚拟化：无需修改OS

**容器** (Docker)：操作系统虚拟化

* + 启动快、资源消耗低
  + 隔离比虚拟机稍弱（共享内核）
  + 组件：Client、Daemon、Image、Container、Registry
  + Kubernetes：开源的容器“操作系统”，负责调度和管理

**分布式系统**

* 定义：由网络连接的多台计算机组成的软件系统
* 特点：可访问性、透明性、开放性、可扩展性

**分布式计算**

* 优点：负载均衡、并行处理、低硬件成本
* 主流技术：
  + 大数据编程：MapReduce、Spark
  + 面向对象：OOM、CORBA、DCOM、Java/RMI
  + 面向服务：SOM、Web Services、CORBAA
  + 共用计算：UBM
  + 志愿计算：VJM
  + 实时大数据：**Storm**

**分布式存储**

* 关键：可靠性
* 包含分布式文件系统（非结构化），分布式数据库系统（结构化）
* 例子：HBase, 构建在分布式文件系统（如 HDFS）之上，适合处理非结构化数据

**分布式文件系统（针对非结构化数据）：**

* NFS：远程访问，不复制文件，本地只读
* Andrew AFS：上传/下载模型
* Kass：基于 Java
* GFS (Google File System)：
  + Master（元数据）、Chunkserver（存储块）、Client
  + Client-Server 系统
* HDFS (Hadoop DFS)：
  + 开源，应用最广
  + NameNode（元数据）、DataNode（数据块）、Client

**分布式数据库系统（结构化数据）**

* 元数据：描述数据属性的信息
* DBMS：包含局部数据库与全局数据库
* 冗余：
  + 分布式 → 增加冗余（但开销高）
  + 集中式 → 减少冗余（但安全性差）
* 数据库结构
  + 全局外层：对外提供服务
  + 全局概念层：整体描述
  + 局部概念层
  + 局部内层
* 数据库一致性ACID 原则
  + A 原子性C 一致性I 隔离性D 持久性
* 分布式系统数据一致性
  + CAP 理论：一致性（C）、可用性（A）、分区容错性（P）；三者只能同时满足 2 个（一致和可用会冲突）
  + BASE 理论：
    - 系统基本可用（Basically Available）
    - 软状态（Soft State，允许暂时不同步）
    - 最终一致性（Eventual Consistency，最终需达成一致）

# 第二章：物联网

**发展时间线**

* 1995：比尔·盖茨《未来之路》首次提及
* 1999：提出“万物互联”，RFID 为重点
* 2003：传感网络提出，但受限于电池问题

**定义**

* 通过感知设备将所有物品连接到物联网，进行信息交换，实现智能化识别与管理
* 即：在互联网的基础上 增加“感知层”

**特征**

* 普通对象设备化，自制终端互联化，普适服务智能化

**MEMS**：

* 微机电系统的缩写，对微米，纳米材料进行设计，加工，制造，测量和控制的技术
* 提高自动化，智能化的可靠水平

**物联网架构（三层 + 平台层）**

* 应用层（含平台层）：提供云服务、计算资源、数据共享与交换
* 网络层：负责数据传输
* 感知层：
  + 感知：传感器、二维码、RFID、多媒体信息
  + 数据采集：短距离组网，协同处理

**感知层技术**

* 智能传感器 VS 标识与自动识别技术
  + **感知信息的内容不同**
* **智能传感器（IEEE1451）**
  + 具备采集、存储、处理、交换、自补偿、自诊断能力；是MEMS和传感器集成化相结合的产物
  + 与传统的传感器区别在于：信号的智能处理
  + 输出信号（电信号）和输入物理量（比如温度）不一定是线性关系
  + 接口：UART（异步，一对一，慢）、SPI（同步，多设备，快）
* **标识与自动识别技术**
  + 信息化的基础工作
  + 包含：自动识别系统，应用程序接口，应用系统软件
  + 条码（1D/二维码；低成本）、IC卡（比SIM卡安全）、**RFID**（非接触式识别，可擦写，EPC global，频率越高传输距离越远，**二代身份证**）、光学字符识别技术OCR、语音识别ASR、生物识别、遥感
* **无线传感网 (WSN)**
  + 由大量传感器组成；和普通传感器不同，具备转发和路由能力
  + 特性：大规模、**自组织网络**、动态、可靠
  + 协议：
    - MAC 协议（**节能**、**时分多址** 避免碰撞；数据链路层）
      * S-MAC：周期性休眠和监听信道是否空闲
      * T-MAC：传感器活跃时间固定
    - 路由协议：能量优先，节点局部视野
      * 能量感知路由：节点主动定期或按需发送数据
      * 查询路由：由 sink 节点发起查询（广播），节点响应
  + **ZigBee** 协议：无线，低速，没有传输层协议，基于 IEEE802.15.4

**应用层通信协议**

* MQTT, COAP, AMQP, STOMP

**物联网通信技术**

* 无线个域网（WPAN）：蓝牙（跳频通信）、TG4（低频率，也叫**ZigBee**，频分多址），TG3（高频率，也叫**UWB**）、红外、NFC（802.15）
* 无线局域网（WLAN）：Wi-Fi（频分多址，802.11）
* 无线城域网（WMAN）：LoRa（802.16）
* 低功耗广域网（LPWAN）：NB-IoT（长距离，**最适合物联网**）
* 移动通信网：2G / 3G / 4G / 5G（基站密度不同，5G距离短，NB-IoT 补充长距离）
* 其他：有线网络、卫星网络
* 例子
  + 互联网 TCP/IP
  + 无线宽带 Wi-Fi
  + 无线低速网 ZigBee，蓝牙

# 第三章：数据基础

**数据交易所**

* 贵阳（第一家），深圳，北京，上海，广州

**大数据的价值**

* 不在于规模，而在于**专业处理**，强调 **真实性（可选）+实时性（必须）**。

**其它**

* 大数据难以找到因果关系
* 数据挖掘属于“二次应用”
* 包结构→完整性；信道编码→正确性
* 数据属性冗余→数据集成；数据冗余→数据消减
* 无线传感网是无中心的，所以抗毁性强
* 四种范式：实验，理论，计算，数据

**大数据处理技术**

* 并行处理数据库，数据挖掘，分布式文件，分布式数据库，云计算平台，互联网和可扩展的存储系统

**大数据处理步骤**

* **数据采集**：数据源、类型、存储与传统不同。
* **数据预处理**（提高质量，减少挖掘时间）：
  + 不完整 → 数据集成（必要）。
  + 不一致 → 数据转换。
  + 噪声 → 数据清洗（必要）。
  + 冗余 → 数据消减、清洗、集成。
  + **传统ETL**：抽取 → 加载 → 转换。
* **数据存储与管理**：
  + **NewSQL**：可扩展，高性能，支持关系模型。
  + **NoSQL**：非关系型，无固定表结构，无连接操作，不严格遵守ACID。
* **数据处理与分析**：MapReduce、Spark。
  + “**分布式并行处理框架**”处理非结构化，半结构化数据
* **数据可视化**

**数据质量**

* **定义层**：如超出范围（86分记成186分）。
* **实例层**：拼写错误、噪声、重复数据（86分记成76分）。
* 分类：
  + **单数据源定义层**：针对单一数据源进行数据结构、字段、类型、业务含义等的标准化定义。
  + **单数据源实例层**：错误值、重复、噪声等。
  + **多数据源定义层**：同一实体不同称呼。
  + **多数据源实例层**：维度、粒度不一致。

**数据分析方法**

* **描述性分析**：发生了什么。
* **诊断性分析**：为什么发生。
* **预测性分析**：可能发生什么。
* **指令性分析**：针对情况应采取什么措施。

**大数据思维**

* **全样而非抽样**：如谷歌流感趋势预测、商品比价网站。
* **效率而非精确**：全样降低误差影响，例：谷歌翻译。
* **相关而非因果**：如纸尿布和啤酒、药品研发。
* **以数据为中心**：搜索引擎、谷歌广告。
* **我为人人，人人为我**：迪士尼手环。

**数据挖掘方法**

* **关联分析法**：**Apriori**、DHP、DIC。
* **数据分割法**：
  + 聚类：数据集没有预定义类别，通过相似性将对象分组。无类别数量。
  + 分类：数据集有给定类别（标签），建立模型对新数据进行分类。有给定类别数量，无法划分的为噪声。
* **偏离分析法**：结果与期望差异。
* **预测模型法**：机器学习。

**大数据分析软件**

* **批处理**：MapReduce、Spark。
* **流计算**：Storm、S4、Puma、Super Mario、Flume。
* **图计算**：Pregel、Hama、GoldenOrb。
* **查询分析**：Dremel、Hive、Cassandra、Impala。

**数据可视化**

* 作用：观测、跟踪、分析、辅助理解、增强吸引力。
* 特性：实时性、简单、丰富展现、多数据集成。
* 工具：
  + 简单：Excel、SQL、Python、R。
  + 专业：Tableau、QlikView、Echarts。

# 第四章：人工智能

**基础认知**

* **反馈型网络**不一定比**前馈型网络**慢。
* **机器学习发展阶段**：共经历 5 个阶段。
* **单层神经网络**：只有两层处理结构（输入层+输出层）。
* **机器学习定义**：通过学习使机器“像人脑一样思考”，但目标不是解决全新问题，而是模拟已有规律。
* **人工神经网络模型**：
  + 前向神经网络
  + 循环神经网络
  + ❌ 没有“后向”和“对称”结构。
* **卷积神经网络（CNN）**：同层之间没有联系。
  + 中间层：卷积层、池化层。
* **循环神经网络（RNN）**：同层之间有联系。
* **深度学习适用场景**：非结构化数据、大量可用数据、足够算力和时间。

**人工智能定义与三要素**

* **定义**：根据环境感知做出合理行动，并获得最大收益的计算机程序。
* **三要素**：**算法、算力、大数据**。
* **实现方法**：编程技术 → 模拟法。

**学习方法**

* **监督学习**：分类、回归。
* **半监督学习**：分类、回归。
* **无监督学习**：聚类、降维。

**学习策略**

* **模拟人脑**：
  + 基于符号（数学公式）。
  + 基于连接（神经网络）。
* **数学方法**：统计学。
* **学习目标**：不强调。

**学习算法**

* **深度学习**。
* **强化学习**。

**机器学习一般流程**

1. 数据获取。
2. 特征抽取（如 n-gram）。
3. 模型训练和验证。
4. 线下测试（训练集验证）、线上测试（真实数据验证）。

**常见机器学习方法**

* **朴素贝叶斯**：
  + 稳定分类效率。
  + 对小规模数据友好。
  + 对缺失数据不敏感。
* **决策树**。
* **支持向量机（SVM）**：
  + 二元分类（是猫还是不是猫）。
  + 多分类需多次 SVM 组合。

**人工神经网络（ANN）**

* **定义**：分布式并行信息处理的数学模型。
* **深度神经网络**：超过三层即为深度神经网络。
* **深度信念网络（DBN）**：基于受限玻尔兹曼机（RBM，双向）。
* **三要素**：神经元特性、拓扑结构、训练规则。
* **连接方式**：
  + 全连接。
  + 局部连接。
  + 稀疏连接。
* **应用**：
  + 卷积神经网络（CNN）：图像分析。
  + 循环神经网络（RNN）：语音识别、语言建模。
  + 深度信念网络（DBN）：语音识别。
  + 生成对抗网络（GAN）：样本生成（由生成模型网络和判别模型网络组成）。

# 错题