第一章 大数据概述

第一章考点

- DIKW模型
- 三次浪潮
- 大数据成因（评估数据治理的情况：数据成熟度（了解））
- 大数据发展历程（了解）
- 大数据特征的定义4V
- 结构化数据和非结构化数据的定义
- 大数据核心技术
- 大数据、云计算与物联网的关系

一、大数据时代

1. 什么是大数据
   大数据是指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系进行记载的物理符号
2. 数据、信息和知识
   数据经过整理和预测变为信息，信息经过提炼和挖掘变为知识
   DIKW模型：数据->信息->知识->智慧
3. 大数据的成因
   三次信息化浪潮：第一次浪潮产生了个人计算机，解决了**信息处理**的问题，第二次浪潮产生了互联网，解决了**信息传输**的问题，第三次浪潮产生了**物联网，云计算和大数据**，解决了**信息爆炸**的问题
   大数据成因：
   - **数据生产方式的变革**促成大数据时代的来临
   - **信息存储技术**为大数据时代提供技术支持
   - **信息处理技术**为大数据技术时代提供技术支撑-CPU处理能力大幅提升
   - **信息的传输技术**为大数据时代提供技术支撑——网络带宽不断增加
4. **大数据的发展历程**
   - 萌芽期与早期探索
     - 1887年-1890年： 美国统计学家赫尔曼·霍尔瑞斯（Herman Hollerith）为了统计1890年的人口普查数据，发明了一台电动器来读取卡片上的洞数。这被视为自动化数据处理的早期雏形。
     - 1944年： 卫斯理大学图书馆员弗莱蒙特·雷德（Fremont Rider）出版了《学者与研究型图书馆的未来》（The Scholar and the Future of the Research Library）一书，预见了大数据时代的到来（指信息爆炸的趋势）。
     - 1997年： 美国宇航局（NASA）研究员迈克尔·考克斯（Michael Cox）和大卫·埃尔斯沃斯（David Ellsworth）首次使用“大数据”（Big Data）这一术语，用来描述

20世纪90年代由超级计算机生成的大量数据信息。

- 技术突破与概念成型
  - 2003年-2006年： Google先后发表了三篇具有里程碑意义的论文，分别介绍了分布式文件系统（GFS）、并行计算模型（MapReduce）和非关系数据存储系统（BigTable）。这三篇论文（常被称为“谷歌三驾马车”）第一次提出了针对大数据分布式处理的可重用方案，奠定了大数据技术的基础。
  - 2008年： 著名科学杂志《Nature》推出了专刊“Big data: The next Google”，首次正式提出“大数据”这一概念，标志着大数据从技术圈走向学术界和公众视野。
- 爆发与普及
  - 2009年： “大数据”逐渐成为互联网信息技术行业的流行词汇。这一时期的研究焦点集中在性能、云计算、大规模数据集并行运算算法以及开源分布式架构（即Hadoop）。
  - 2012年1月： 世界经济论坛在瑞士达沃斯召开，大数据成为主题之一。会上发布的报告《大数据，大影响》宣称：“数据已经成为一种新的经济资产类别”。
  - 2013年： 大数据技术开始向商业、科技、医疗、政府、教育、经济、交通、物流等多领域全面渗透。因此，2013年也被称为“大数据元年”。**

二、大数据概念

大数据特征4V

- **数据量大(Volume)**
- **数据类型繁多(Variety)**：**结构化数据和非结构化数据**，其中结构化数据是行数据，可以用二维表来逻辑表达实现的数据，非结构化数据则是不能用二维表来逻辑表达实现的数据
- **处理速度快(Velocity)**：从生成到消耗，时间窗口非常小，可以用于生成决策的时间非常少
- **价值密度低(Value)**：价值密度低，商业价值高

三、大数据的影响及应用

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**：无需抽样，直接考虑整体
- **相关而非因果**：相关并不代表因果
  在社会发展方面：大数据决策逐渐成为一种**新的决策方式**
- 大数据应用有力地**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

四、大数据关键技术

1. 大数据技术的不同层面
   - 数据采集
   - 数据存储和管理

- 数据处理与分析
- 数据隐私和安全

2. 大数据的**两大核心技术**
   - **分布式存储**
   - **分布式处理**

五、大数据产业

大数据产业是指一切与支持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产业
数据源层	大数据生态圈有各种提供来源
数据平台层	数据分享，数据分析，数据租售等
数据应用层	提供智能服务

六、大数据与其他新兴技术

1. 云计算
   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要的各种IT资源**
   云计算的服务模式和类型主要包含这三类：
   - 软件即服务(SaaS)
   - 平台即服务(PaaS)
   - 基础设施即服务(IaaS)
     云计算的关键技术：
   - **分布式存储**：实现大规模数据的高可靠存储，是云计算的数据基础。
   - **虚拟化**：通过资源虚拟化提高资源利用率和灵活性，是云计算的核心支撑技术。
   - **分布式计算**：实现大规模任务的并行处理，是云计算的计算基础。
   - **多租户**：允许多个用户共享同一物理资源，同时保障数据隔离和安全，是云服务的关键特征之一。
2. 物联网
   物联网是物物相连的互联网，是互联网的延伸，它利用局部网络或者互联网等通信技术奖传感器、控制器、机器、人员和物等通过新的方式连在一起，形成人与物、物与物相连，**实现信息化与远程管理控制**
3. **大数据、云计算与物联网的关系**
   大数据、云计算和物联网代表了IT领域最新的技术发展趋势，三者既有区别又有联系。
   三者侧重点：
   - 数据侧重于对海量数据存储、处理与分析，从海量数据中发现价值，服务于生产和生活。

- 云计算旨在整合和优化各种IT资源并通过网络以服务的方式，廉价地提供给用户。
- 物联网的发展目标是实现“物物相连”，应用创新是物联网发展的核心
  三者的区别与联系
- 云计算为大数据提供了技术基础，大数据为云计算提供了用武之地
- 物联网是大数据的重要来源，大数据技术为物联网数据分析提供支撑
- 云计算为物联网提供海量数据存储能力，物联网为云计算提供了广泛的应用空间

4. 大数据与人工智能
   - 联系：AI需要数据来建立其智能，特别是机器学习。大数据技术为人工智能提供 了强大的存储能力和计算能力。
   - 区别：I是一种计算形式，它允许计算机执行认知功能，例如，对输入起作用或 做出反应，而大数据技术只是寻找结果。大数据主要通过对比分析来推演出更 优的方案，而人工智能是为了辅助或代替我们更好地完成某些任务