复习的内容:课本第1，2，3，4，5，6，7童

一、大数据基础概念

1.大数据的定义与特征

定义:无法用常规工具处理的海量、高增长、多样化数据集合。

5V特征:volume(大量)、Velocity(高速)、variety(多样)、Value(低价值密度)、Veracity(真实性)。“

2、数据结构类型

结构化数据:如关系型数据库表格(学生成绩表、销售数据)。

半结构化数据:如 XI/JSON 配置文件、日志文件、

非结构化数据:如图片、视频、音频、文本(微博内容、日志)。

二、大数据技术体系

(一)数据采集与预处理

1.采集技术

科研数据采集:大型强子对撞机、射电望远镜、电子显微镜。

互联网数据采集;网络爬虫、传感器节点、社交媒体服务器。

2.预处理方法

数据清洗：忽略元组、人工填充缺失值、分箱法。

数据集成：模式匹配、实体识别、数据聚合。

数据转换与归约：标准化、归一化、特征选择。“

(二)数据存储与管理“

1.存储架构

DAS(直接附加存储):直接连接服务器，适合小规模场景，扩展性差。

NAS(网络附加存储):通过网络共享文件，支持多用户，性能受带宽限制。-

SAN(存储区域网络):专用光纤网络，块级访问，高带宽低延退，本高。



2.存储技术

分布式文件系统:HDFS(Hadoop核心)、云存储(Amazon s3、阿里云OSS)。

NoSQL 数据库:MongoDB(文档型)、Cassandra(列族)，支持高扩展非结构化数据。

关系型数据库:MySQL、PostgreSQl，保证数据一致性。

NewSQL:结合关系型与 NoSQL 优点，支持 ACID 事务。

(三)数据处理与分析框架

1、Hadoopf

核心组件:HDFS(分布式存储)、MapReduce(批处理计算)、YARN(资源调度)

适合大规模批处理任务，对硬件要求低，但磁盘I0 开销大。

2、Spark

核心组件:Spark Core(内存计算)、Spark SQl、Spark Stxeaming、MLlib(机器学习)、graphX(图计算)。

基于内存计算，处理速度快，适合实时计算和迭代任务(如聚类、关联规则）。

1. 其他技术

云计算服务模式:Iaas(基础设施)、Paa8(平台)、Saas(软件)，Daas(数据即服务)不属于核心模式。

(四)数据分析与挖掘算法

1.分类算法:决策树、朴素贝叶斯、支持向量机(SVM)、贝叶斯网络。。

2.聚类算法:K-Means、DBSCAN。2.

3.关联规则算法:Apriori:、FP-growth(扫描次数少，基于 Apriori优化)。

4、其他任务:回归分析、特征选择(减少维度，增强理解)。

三、大数据应用场景

1.金融领域:欺诈检测、信用评估、股票预测。-

2.医疗领域:疾病预测(如流感流行分析)、精准医疗(癌症个性化治疗)、医疗影像诊断、药物研发。

3.交通与物联网:智能交通管理、自动驾驶(边缘计算应用)。

4.社交与互联网:社交网络分析、用户行为分析(Hadoop处理海量曰志)。-

5.科研与企业:科研数据处理(天文观测、粒子对撞)、企业 BI(商业智能，如Tableau 分析销售数据)。

四、基础设施与家全

1.数据中心

分级:IV级可用性最高，具备高可靠性、可扩展性、绿色环保。

网络架构：核心层，汇聚层，接入层

1. 安全与隐私

挑战：数据泄露，隐私侵犯（如医疗数据脱敏），完整性威胁，访问控制复杂

应对：加密，脱敏，访问控制，安全审计

1. 新兴技术与发展趋势
2. 未来前景

应用扩展：农业，能源，远程医疗等邻域渗透

技术融合：与AI，物联网结合（如智能设备数据实时分析）

1. 瓶颈问题

数据质量：缺失，错误数据影响分析

安全隐私：数据泄露风险

人才与成本：专业人才稀缺，计算资源（硬件，能耗）成本高

模型局限性：复杂模型解释性差，稳定性不足

六、常见误区与概念辨析

1.大数据≠PB 级数据:TB 级数据即可视为大数据，关键在于处理模式。“

2.Hadoop与Spark 的区别:Hadoop 适合批处理，Spark 适合实时和迭代计算。

3.数据仓库(DW)与数据库(DB):DW用于OLAP(分析型)，DB用于0LTP(事务型)，不可互换。

4.NoSQL与关系型数据库:NoSQL不替代关系型数据库，二者适用场景不同(如事务性需求仍依赖关系型数据库)。

七、简答题

1.大数据是什么?它有哪些特点?

大数据是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合,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能力的海量、高增长率和多样化的信息资产.其特点通常用5V来概括:Volume(大量)，数据规模庞大从TB级别跃升到PB级别甚至更高:Velocity(高速)，数据产生和处理速度快，需要实时或准实时处理;Variety(多样)，数据类型繁多，包括结构化、半结构化和非结构化数据，如文本、图像、音频、视频等:Value(低价值密度)，在海量数据中，有价值的信息密度相对较低，Veracity(真实性)强调数据的准确性和可靠性，数据来源复杂，需保证数据质量。

2.大数据有哪些结构类型?每种数据有什么特点?请分别举例。

大数据结构类型主要有结构化、半结构化和非结构化数据。结构化数据特点是数据具有固定的格式和模式，以表格形式组织，易于查询和管理，例如关系型数据库中的用户信息表，包含姓名、年龄、性别等字段，每一行记录对应一个用户的信息;半结构化数据有一定的结构，但结构不严格，通常以自描述的方式存在，比如XM文件，它有标来标识数据元素，但不像关系型数据库那样有严格的表结构，非结构化数据没有预定义的结构，数据格式不规则，例如文本文件(小说、论文等)、图像文件(照片、设计图等)、音频文件(音乐、语音等)

3.请详细比较 DAS、NAS 和 SAN 三种存储架构的特点。

DAS(Direct-Attached Storage,直接附加存储)是将存储设备直接连接到服务器，特点是连接简单，数据传输速度快，适合小规模、对性能要求高且数据量相对固定的场景，但缺点是可扩展性差，存储设备与服务器紧密绑定，资源利用率低，管理分散。NAS(Network-Attached Storage，网络附加存储)通过网络连接存储设备和服务器，基于文件系统进行数据存储和访问，特点是部署方便，可通过网络远程访问，支持多用户共享，易于管理;但由于基于网络传输，在高并发大流量访问时，性能可能受网络宽带限制。SAN（Storage Area Network，存储区域网络)是通过专用能可能受网络带宽限制。SAN(Storage Area Network,光纤网络连接存储设备和服务器，以块级方式访问数据，特点是具有高带宽、低延迟，数据传输性能强，可扩展性好，能集中管理存储资源，但缺点是建设成本高，需要专业的光纤网络设备和管理技术。

4.请分析大数据分析未来的发展前景以及未来可能会遇到的瓶颈和问题。

操作方面可以学习的内容:

黑马程序员 Python 教程，4 天快速入门 Python 数据挖掘，系统精讲-实战案例

观总税净员2Y1AW服标，4队速入广T技狐控锋，