**本文档收集整理了大数据技术导论一书中的习题，并对习题通过AI利用教师提供的英文PPT，乐学平台发布的中文PPT结合个人理解进行了回答，仅供参考+1.**

**其中注明出现的专有名词，部分来源于维基百科，仅供参考+2.**

**标黄处是PPT中出现，以及个人觉得可能考察的题目，仅供参考+3.**

**至于最后考试会不会有用，仅供参考+4.**

**习题一**

1. **云计算与大数据存在何种关系?**
2. **中国在大数据领域较发达国家有哪些不足?（这题能考我吃）**
3. **为什么说人们进入了大数据时代?**
4. **什么是分布式存储?**
5. **大数据技术相较于传统技术有哪些不同?**
6. **大数据有哪些特征?**
7. **面对大数据密度价值低的特性，如何规避随少量有效数据而来的大量无效数据的干扰?**
8. **哪些常用的软件会产生大数据?**

回答：

1. 云计算 cloud computing

云计算提供了大数据存储和处理所需的基础设施和弹性计算资源，而大数据则需要云平台提供的高效、可扩展的环境来处理和分析海量数据。

具体而言，

1）云计算提供大数据所需的基础设施大数据指的是超出传统数据处理能力的大规模数据集合，这些数据量庞大、类型多样、增长迅速。大数据的处理需要强大的存储和计算能力，而这些正是云计算的优势所在。云计算提供了：弹性存储：云平台提供可扩展、灵活的存储解决方案，可以根据大数据的规模变化动态扩展存储空间。例如，云中的对象存储（如AWS S3、Azure Blob Storage）能有效地存储海量结构化和非结构化数据。弹性计算：云平台提供按需分配的计算资源，允许用户根据数据处理需求动态增加或减少计算能力。例如，计算集群可以通过云计算平台（如AWS EC2、Google Cloud Compute Engine）进行快速扩展，满足大数据分析时对计算资源的需求。（2） 云计算支撑大数据的分布式处理大数据处理通常需要进行分布式计算，而云计算本身具备分布式架构的特点。云计算通过虚拟化技术和分布式计算资源，能够提供大数据处理所需的计算能力，支持并行处理和分布式存储。例如，常见的大数据处理框架（如Hadoop、Spark）可以部署在云计算平台上，通过多台计算机的协同工作实现大规模数据的高效处理。Hadoop和Spark框架：云平台（如AWS、Google Cloud、Azure）通常提供与大数据框架的集成支持，允许用户在云环境中部署Hadoop、Spark等框架进行大数据处理。这些框架利用云的分布式计算资源，支持批量数据处理和实时流处理。分布式数据库和NoSQL数据库：如HBase、Cassandra等可以在云环境中部署，支撑对大数据的高效存储和查询。（3.）云计算提供大数据分析和机器学习平台大数据不仅仅是存储和处理问题，更多的是如何从海量数据中提取有价值的信息。云计算平台提供了强大的分析和机器学习工具，帮助企业从大数据中获得洞察。数据分析工具：如Google BigQuery、Amazon Redshift、Azure Synapse等云数据仓库，提供了大数据查询、分析和报告功能，支持SQL、NoSQL等多种查询方式。机器学习平台：云计算提供了机器学习（ML）和人工智能（AI）服务，如AWS SageMaker、Google AI Platform、Azure Machine Learning等，允许用户基于大数据进行训练模型、预测分析等任务。这些平台提供了丰富的算法库、自动化的机器学习工具和高性能的计算资源。（4.）云计算与大数据的成本效益大数据的存储和处理往往涉及巨大的成本，尤其是在需要处理海量数据时。传统的IT基础设施需要大量的投资和运维，而云计算通过按需付费的方式，降低了企业获取计算和存储资源的门槛，提供了显著的成本效益。按需付费：用户可以根据使用量支付费用，而无需预先投资大量硬件设备。随着数据量和计算需求的增加，云计算可以动态扩展资源，避免了过度采购的浪费。减少运维成本：云平台提供自动化的资源管理和监控工具，减少了对专业IT运维人员的依赖。（5.）云计算支持大数据的实时性和可扩展性随着大数据技术的发展，实时数据处理和分析变得越来越重要。云计算能够提供高可扩展性和低延迟的数据处理能力，支持大数据的实时流处理和实时分析。实时数据流处理：云平台（如AWS Kinesis、Google Cloud Dataflow、Azure Stream Analytics）可以处理实时数据流，为企业提供实时监控、分析和决策支持。高可用性和容错性：云计算平台通常提供高可用性设计，能够确保大数据处理任务即使在某些计算节点或存储节点发生故障时，仍然能够继续运行，保障了系统的稳定性和可靠性。（PPT上没有的东西你要考这么细我吃）

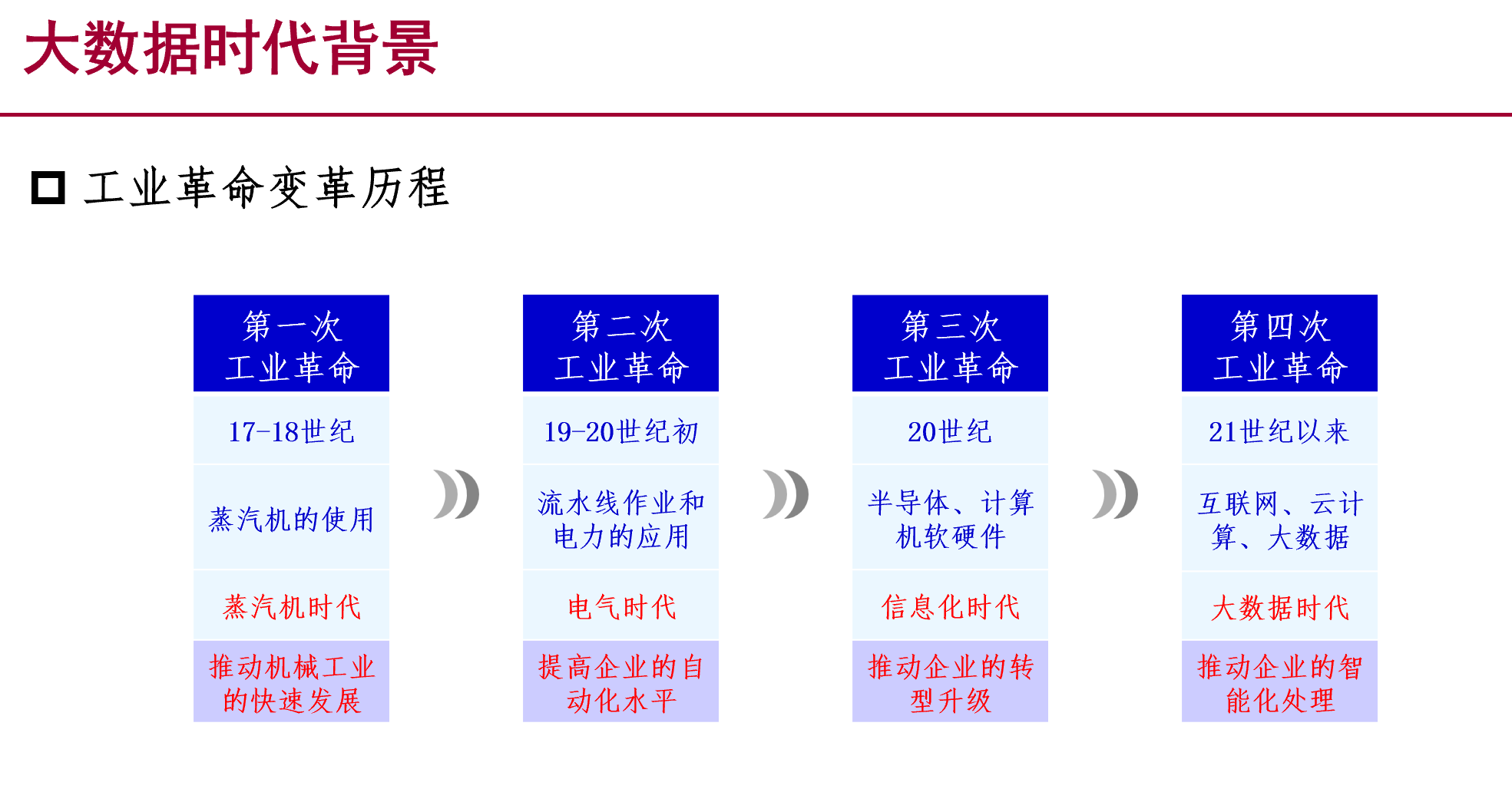
2.

数据隐私与法律法规：虽然中国有《个人信息保护法》，但在数据隐私保护和跨境数据流动的法规执行方面，相较欧美国家仍显滞后。欧洲的GDPR等法律体系更为严格。核心技术与创新：中国在大数据的基础技术（如分布式计算、深度学习算法）上的创新仍不如美国等发达国家，主要依赖国外技术。数据共享与合作：中国存在“数据孤岛”问题，行业间和跨境数据共享不足，数据流动受限。而欧美在数据共享机制和跨行业合作方面已较为成熟。应用深度与产业整合：在工业互联网、精准医疗等领域的应用较为初步，整体产业整合程度不高。发达国家的应用更具深度和广度。云计算与服务生态：中国的云计算基础设施发展较快，但在全球市场和服务生态上仍落后于AWS、Azure等国际巨头。人才和教育体系：高端数据科学人才流失问题较为严重，且学术界与产业界的合作不够紧密，人才培养和教育体系与实际需求有一定脱节。标准化不足：大数据领域的技术标准化工作起步晚，缺乏统一标准，影响了技术的兼容性和应用的广泛性。（能考这个我吃好吧）

3.

人们进入大数据时代是因为数据的生成、存储和分析能力大幅提升。现代技术使得海量数据（如社交媒体、传感器、交易记录等）快速产生和收集，超出了传统数据处理方式的能力。随着云计算、人工智能、物联网等技术的成熟，数据可以被实时分析和应用，推动了各行各业的变革。大数据的价值在于其深度分析能够揭示规律、预测趋势，带来决策优化和创新机会

（PPT中下图好像涉及了这一点）



4.（感觉这题蛮重要，PPT课后题也有这个）

分布式存储是一种将数据分散存储在多台计算机或存储设备上的技术。这些存储设备通过网络连接，在一个集群中共同协作，实现数据的存储、管理和访问。与传统的集中式存储（将数据存储在单个服务器上）不同，分布式存储通过多个节点（即服务器）分担数据存储的任务，从而提高了系统的可靠性、扩展性和容错能力。

总的来说，分布式存储通过集群化的方式提升了数据的存储能力、处理能力和可用性，满足了现代大数据、云计算等应用对海量数据存储的需求。

5.

1. 数据规模（Volume）

* 传统技术：传统数据处理技术通常针对结构化数据，数据规模相对较小，处理的数据量一般在TB级别以下。数据存储和处理多依赖于单一的数据库系统，扩展性有限。
* 大数据技术：大数据技术则可以处理海量数据，通常在PB级别甚至更大。它能够存储和分析来自不同来源的数据，包括结构化数据、半结构化数据（如日志、XML文件）和非结构化数据（如文本、图像、视频等）。

2. 数据类型与结构

* 传统技术：传统技术主要处理结构化数据，数据通常存储在关系型数据库中，数据模型固定（如表格形式），需要在预定义的模式下进行存储和查询。
* 大数据技术：大数据技术不仅能处理结构化数据，还能处理半结构化数据（如JSON、XML格式）和非结构化数据（如文本、音视频文件等）。这要求数据存储和分析技术更加灵活，可以支持多种数据格式。

3. 处理方式（批处理 vs 实时处理）

* 传统技术：传统技术主要依靠批处理方式，数据处理通常是离线进行的，处理周期较长。数据分析通常是周期性进行，如每日、每周等。
* 大数据技术：大数据技术不仅支持批处理，还能进行实时数据处理。通过流处理框架（如Apache Kafka、Apache Flink等），大数据系统可以实时收集、分析和响应数据，适应快速变化的环境。

4. 存储方式

* 传统技术：传统技术通常使用单机存储或关系型数据库来存储数据，数据存储方式相对单一，扩展性差。扩展存储能力时，通常需要购买更高性能的硬件或通过垂直扩展提升单机性能。
* 大数据技术：大数据采用分布式存储技术（如Hadoop HDFS、Ceph等），数据分布在多个存储节点上，具备更好的扩展性和容错能力。数据可以跨多个节点存储，并且通过冗余和备份技术提高可靠性。

5. 计算框架与处理方式

* 传统技术：传统技术通常使用单机计算，数据处理多依赖于关系型数据库管理系统（RDBMS）或小规模的数据处理框架。计算能力受限于单机的硬件性能。
* 大数据技术：大数据技术采用分布式计算框架（如Hadoop、Apache Spark等），能够将数据分布到多个计算节点上进行并行计算，显著提高了处理速度和计算能力，适应了大规模数据的需求。

6. 分析方法

* 传统技术：传统分析方法多侧重于历史数据分析，采用SQL查询语言进行数据筛选、排序、聚合等操作。通常分析方式较为简单，无法处理复杂的数据模式。
* 大数据技术：大数据技术支持更加复杂和多样化的分析方法，如机器学习、深度学习、图计算、时序分析等。大数据不仅能处理静态数据，还能实时分析、预测未来趋势，支持更复杂的决策和洞察。

7. 灵活性与可扩展性

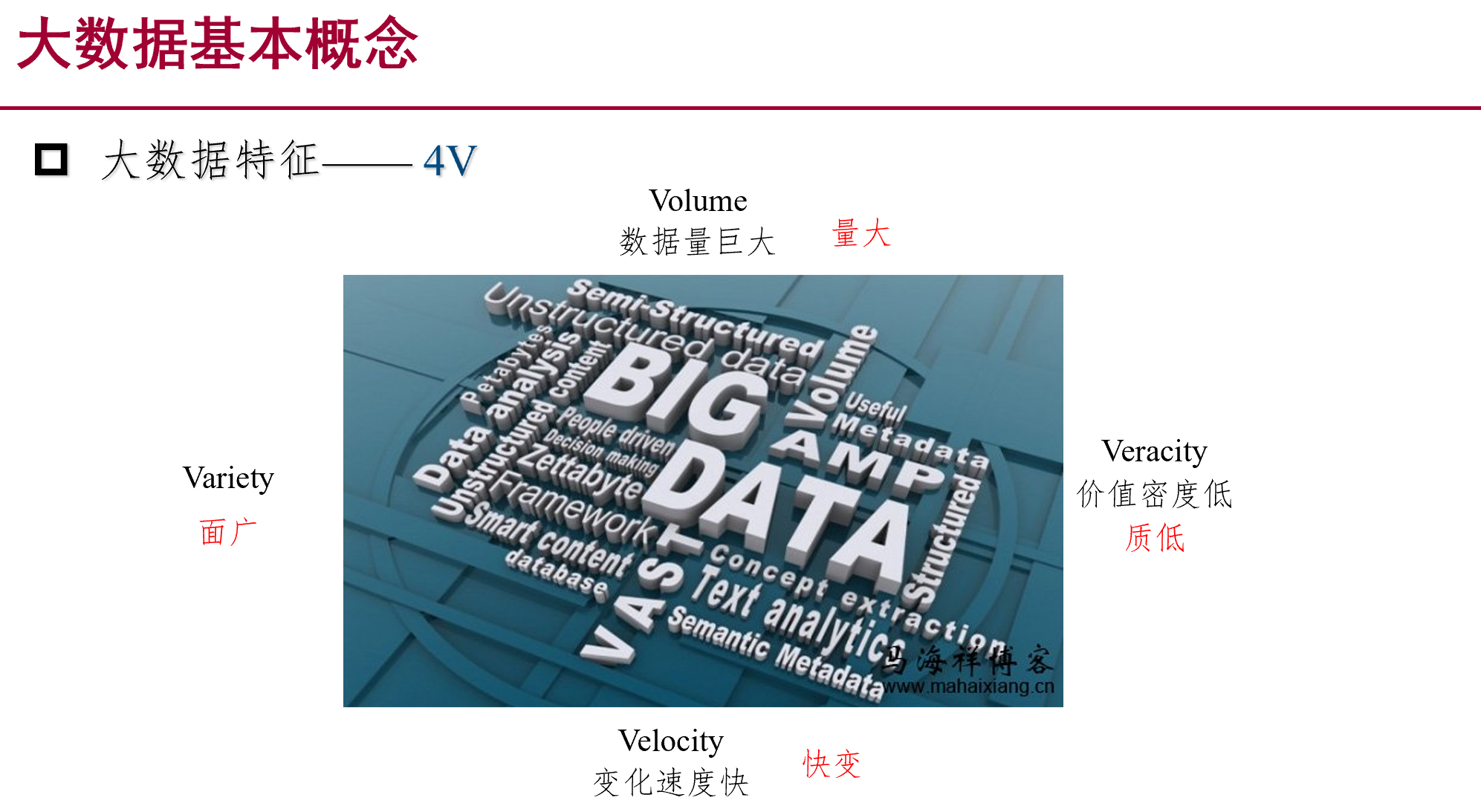
* 传统技术：传统系统的扩展性相对较差，通常需要通过增加单机硬件的性能（如增加CPU、内存、硬盘等）来应对数据量的增长，难以快速适应不断变化的需求。
* 大数据技术：大数据系统具备高度的可扩展性，通过横向扩展（增加更多的存储和计算节点）来应对数据增长，能够在不同规模下灵活应对大量数据的处理需求。

8. 容错与高可用性

* 传统技术：传统技术在单机系统中运行时，往往缺乏有效的容错机制，一旦出现硬件故障或数据损坏，系统容易发生崩溃或数据丢失。
* 大数据技术：大数据技术通过分布式架构和数据冗余技术确保系统高可用性，多个副本存储在不同节点，即使某些节点发生故障，系统也能够继续正常工作，保证数据不丢失。



6. 见下图

. 

7.

1. 数据清洗（Data Cleaning）

数据清洗是去除无效数据、修复错误数据的关键步骤，目的是提升数据的质量，确保后续分析能够依赖有效数据。数据清洗包括：

* 去重：删除重复记录，避免冗余数据干扰分析。
* 缺失值处理：填补或删除缺失的数据项。
* 异常值检测：识别并去除那些极端、不合常理的异常数据点，减少其对模型的影响。
* 格式统一：将数据转化为统一的格式，避免因格式不一致造成的分析困难。

2. 数据筛选与特征选择

在大数据分析中，往往需要从大量的数据中筛选出与目标最相关的数据。这可以通过以下方法来完成：

* 特征选择：通过统计方法（如卡方检验、信息增益、相关系数等）选取与预测任务最相关的特征，去除无关或冗余的特征。
* 降维技术：如主成分分析（PCA）、线性判别分析（LDA）等，通过减少数据维度，保留对模型训练最有价值的信息，从而降低无效数据的干扰。
* 过滤器方法：根据业务需求或数据特性，设定一定的阈值，筛选出数据中的有效部分，剔除不符合要求的数据。

3. 数据采样

在面对海量数据时，不必每次都使用全部数据进行处理，可以通过数据采样来减少无效数据的影响：

* 随机采样：从数据集中随机选择一部分数据进行分析，以减少无效数据的比例。
* 重要性采样：根据数据点的重要性或权重进行采样，只选择对结果影响较大的数据。
* 分层采样：根据数据的不同类别或特征进行分层，从每个类别中按比例抽取数据，保证样本的代表性。

4. 噪声过滤与模型优化

无效数据通常会被视为“噪声”，影响数据模型的训练效果。以下方法可以有效降低噪声的干扰：

* 噪声过滤：使用专门的算法（如低通滤波、SVM中的核方法等）去除数据中的噪声，提升模型的鲁棒性。
* 稳健模型：使用能够处理噪声和异常值的算法，如决策树、随机森林、稳健回归等，这些算法能自动减弱无效数据的影响。
* 异常检测：在数据预处理阶段，使用异常检测算法（如孤立森林、DBSCAN聚类等）识别并剔除异常值，从而提高数据集的质量。

5. 数据融合与多源数据整合

大数据通常来自不同的来源，这些数据可能存在质量差异。通过数据融合与整合，可以提升有效数据的质量，降低无效数据的影响：

* 数据合并：整合来自不同来源的数据，去除重复、无关的信息，增强数据的准确性和可靠性。
* 多视角分析：结合多种数据来源（如传感器数据、日志数据、社交媒体数据等），通过交叉验证和整合，提升数据的有效性。

6. 基于业务场景的智能数据过滤

针对具体业务场景，可以设计智能数据过滤机制，确保只有与目标任务密切相关的数据才被采集和处理：

* 领域专家知识：在数据采集和预处理阶段，结合行业专家的知识，设定数据筛选标准，避免无效数据的干扰。
* 动态数据选择：通过实时分析和动态选择，确保仅将对当前分析任务有价值的数据输入到模型中。

7. 模型训练中的正则化技术

在构建机器学习模型时，正则化技术能够有效避免无效数据和噪声的干扰：

* L1正则化（Lasso）：通过对模型参数的惩罚，促使模型选择更少、更重要的特征，减少无效数据的影响。
* L2正则化（Ridge）：对模型参数进行约束，避免过拟合，减少数据噪声对模型的影响。

8.

1. 社交媒体平台（提到）

社交媒体平台生成的大数据主要包括用户的互动行为、发布的内容、评论、点赞、分享等。这些平台通常拥有成千上万的活跃用户，数据量庞大。

* Facebook、Twitter、Instagram：这些平台每天都产生海量的用户行为数据，如帖子、评论、图片、视频、互动、地理位置等。
* TikTok、YouTube：视频平台生成的视频观看数据、评论、点赞、分享等也是大数据的一部分。

2. 电子商务平台（提到）

电子商务平台生成的数据包括用户浏览行为、购买记录、商品评价、支付信息、物流追踪等。

* Amazon、淘宝、京东：这些平台在用户浏览商品、下单、支付、配送等各个环节生成海量数据，涵盖商品、价格、用户行为、交易信息等。
* eBay、Shopify：这些平台也产生大量的交易数据、用户行为数据和库存数据。

3. 金融软件与银行系统

金融行业涉及到大量的交易、账户、市场数据等，生成的交易记录、市场波动、用户账户信息等都属于大数据的一部分。

* 支付宝、微信支付、PayPal、银行交易系统：这些平台生成的用户交易数据、支付行为、资金流动、信用记录等，都是大数据的典型来源。
* 股市交易系统、基金管理软件：股市交易软件和相关金融分析工具每天都产生大量的交易数据、市场行情、证券价格波动等。

4. 物联网（IoT）平台（提到）

物联网设备生成的数据量巨大，涉及到传感器、设备状态、实时监测等信息。

* 智能家居设备：如智能温控、智能安防、智能电器等设备每天都生成大量的传感器数据、设备使用日志等。
* 工业物联网：工业设备、传感器、生产线监控系统等生成的生产数据、设备运行状态、温度、湿度、压力等也构成大数据的一部分。

5. 搜索引擎与广告平台（提到）

搜索引擎和广告平台每天都产生大量的用户查询、点击、广告展示和交互数据。

* Google、Bing、百度：用户的搜索行为、点击数据、网页访问量、广告互动等，都属于大数据的范畴。
* 广告平台（如Google Ads、Facebook Ads）：广告平台产生的大量广告展示、点击、转化率等数据，也是大数据的来源。

6. 在线游戏平台

在线游戏产生的用户行为数据、游戏进度、虚拟物品交易等，都是大数据的一部分。

* Steam、Epic Games、腾讯游戏：这些平台记录玩家的游戏行为、进度、排名、社交互动、虚拟货币交易等。
* 手游（如王者荣耀、PUBG等）：手游生成的实时玩家行为数据、游戏状态、玩家对战记录等也是大数据的来源。

7. 健康与医疗软件

医疗健康行业的数字化转型使得医疗数据量暴增，包括电子病历、患者监控数据、基因组数据等。

* 电子病历系统（EMR）：医院和诊所生成的病历数据、诊疗记录、实验室检测数据等。
* 健身追踪应用（如Fitbit、Apple Health、运动健康App）：这些软件生成的用户健康数据，包括运动量、睡眠数据、心率等。
* 基因组数据：基因组学领域生成的大量基因数据，如基因组测序数据，属于典型的大数据。

8. 交通与物流管理软件

交通管理和物流系统生成的数据也非常庞大，涉及到位置、路线、交通流量等信息。

* Uber、滴滴、Lyft：这些共享出行平台生成的实时位置数据、用户行程数据、路线优化、司机信息等，都是大数据的来源。
* 物流与供应链管理系统（如FedEx、DHL）：物流平台跟踪运输信息、货物位置、运输时效等数据。

9. 智能制造与工业4.0

智能制造领域的传感器、生产设备、质量监控等系统产生的数据量巨大。

* 工业自动化系统：如PLC控制系统、SCADA系统等，实时监控设备的运行状态、温度、压力、生产速度等数据。
* 工业物联网（IIoT）：生成的数据包括机器设备状态、生产线数据、维护记录等。

10. 内容管理与社交协作工具

在线协作工具和内容管理平台产生的数据主要是文档、项目管理、协作记录等。

* Google Docs、Microsoft Teams、Slack：这些平台记录用户的实时协作、文件更新、消息传递、项目进度等。

11. 传感器与监控系统（提到）

监控视频、传感器、卫星数据等系统生成的数据也属于大数据的一部分，特别是在安全和环境监控方面。

* 视频监控系统：产生大量的图像和视频数据，尤其是在安防、交通监控等领域。
* 气象数据、环境监测：气象站、环境传感器等系统生成的气温、湿度、污染物浓度等数据。

**习题二**

1. **大数据主要有哪些来源?**
2. **如何实现数据的价值?**
3. **什么是组织内部数据?组织内部数据包含哪些内容?**
4. **如何实现企业内部数据价值?**
5. **有哪些常用的组织内部数据获取方法?**
6. **什么是数据抽取?有哪些常用的数据抽取方法?**
7. **什么是数据转换?**
8. **什么是组织外部数据?组织外部数据包含哪些内容?**
9. **组织外部数据按照获取方法可以分为哪两类?**
10. **浅网数据获取方法是什么?**
11. **网络爬虫爬取流程是什么?**
12. **有哪些常见的网络爬虫策略?**
13. **什么是深网数据?如何获取深网数据?**

回答：

1.



 互联网数据

* 社交媒体数据：例如Twitter、Facebook、微博等平台上的用户动态、评论、分享、点赞等行为数据。
* 搜索引擎数据：如Google、百度等搜索引擎中用户的搜索查询记录，反映了用户的兴趣和需求。
* 电子商务数据：如淘宝、京东、亚马逊等平台的购物行为数据，包括浏览记录、购买记录、用户评价等。

 传感器数据  
随着物联网（IoT）技术的发展，许多设备和传感器（如智能家居设备、工业机器、智能城市基础设施等）会不断生成数据。例如：温湿度传感器、GPS定位数据、智能电表数据、汽车传感器数据等。

 企业业务系统数据  
企业在日常运营过程中会生成大量的数据，如客户关系管理（CRM）系统、企业资源规划（ERP）系统、财务管理系统等，这些都包含了客户信息、销售数据、库存数据等。

 移动设备数据  
智能手机、平板电脑、可穿戴设备等产生的日志和传感器数据，包括位置数据、使用行为数据、健康数据等。

 公共数据集和政府数据  
政府和公共机构发布的开放数据（如天气数据、交通流量、人口统计数据等）也是大数据的重要来源。开放数据可以用于研究、决策支持等多个领域。

 日志数据  
各种系统、应用和网络产生的日志文件，记录了用户行为、系统性能、错误报告等信息。这些日志数据通常非常庞大，需要经过清洗和分析才能提供有价值的洞察。

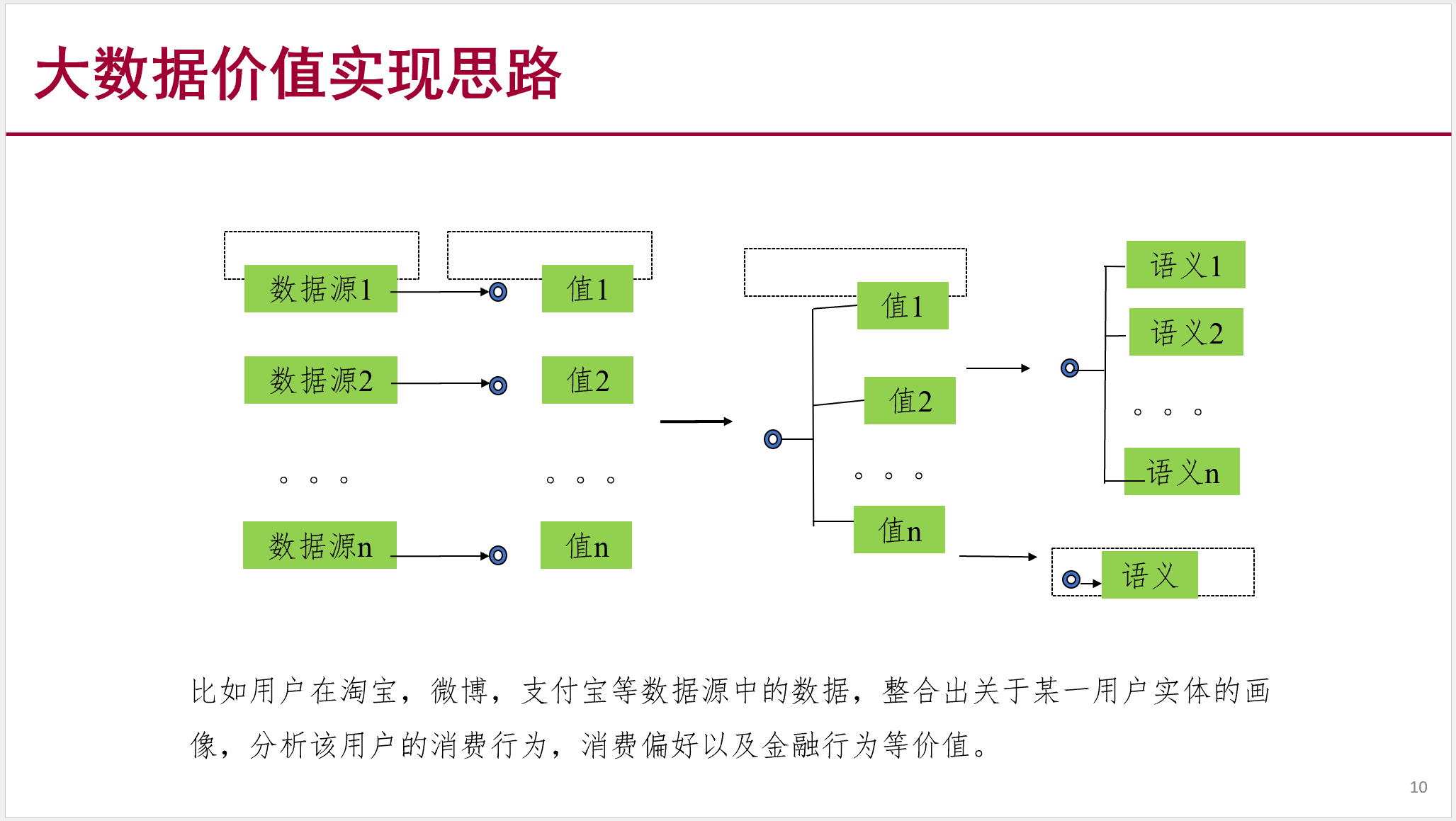
 媒体和娱乐数据  
包括电影、音乐、新闻、视频流平台等产生的数据，例如Netflix、YouTube、Spotify等平台上的用户观看和收听习惯。

 金融数据  
股票市场、银行交易、保险公司记录、支付平台等产生的大量数据。这些数据对于分析市场趋势、投资决策等非常重要。

 医学健康数据  
电子健康记录（EHR）、基因组学数据、医疗影像数据、患者监测设备等生成的大量数据，能够用于疾病预防、治疗效果评估和医学研究。

 文本和语音数据  
来自新闻、书籍、社交媒体、客服对话、电话语音等的文本或语音数据。通过自然语言处理（NLP）技术分析这些数据可以提取出有价值的信息。

（感觉有点像习题一的最后一题）



（我请问这个PPT说了什么。。。）

GPT version

1. 数据收集与获取

* 多渠道收集：确保从传感器、日志、社交媒体、企业应用等多个来源获取数据。
* 数据质量管理：确保数据的准确性、完整性和一致性，为后续分析奠定基础。

2. 数据存储与管理

* 选择合适存储技术：采用如Hadoop、NoSQL等适应大数据量存储的技术。
* 数据清洗与预处理：去除无关、错误、重复的数据，填补缺失值，确保数据质量。
* 数据整合：将不同来源、格式的数据进行标准化和融合，以便统一分析。

3. 数据分析

* 描述性分析：通过统计和可视化，了解数据的基本模式和趋势。
* 诊断性分析：分析数据中隐藏的原因，揭示事件发生的背后原因。
* 预测性分析：利用机器学习等技术预测未来的趋势和事件。
* 规范性分析：根据分析结果，给出优化和决策建议。

4. 数据挖掘与机器学习

* 数据挖掘：使用算法（如聚类、关联分析）发现数据中的模式和规律。
* 机器学习：基于历史数据建立模型，并进行预测或分类应用，如推荐系统和风险预测。
* 深度学习：对复杂数据（如图像、语音）进行更深入的分析，应用于自动驾驶、语音识别等领域。

5. 数据可视化

* 图表展示：通过直观的图形（如柱状图、折线图）帮助理解数据和趋势。
* 交互式可视化：使用工具如Power BI、Tableau，让用户根据需求深入探索数据。
* 地理信息可视化：用于展示空间数据，如位置分析、流量监控等。

6. 数据驱动决策

* 基于数据做决策：将数据分析结果转化为具体的决策，帮助提高效率和精确度。
* 实时决策：在某些应用场景中，如金融交易或智能交通，快速的实时数据分析能产生显著价值。

7. 数据产品化与创新

* 构建数据产品：如推荐系统、智能分析工具等，基于数据分析成果推出具体产品。
* 创新应用：推动新的商业模式和服务创新，如共享经济平台、智能零售等。

8. 数据隐私与安全

* 隐私保护：采取加密、脱敏等措施保护用户隐私。
* 数据安全：确保数据在存储、传输、分析过程中的安全性，防止泄漏和滥用。

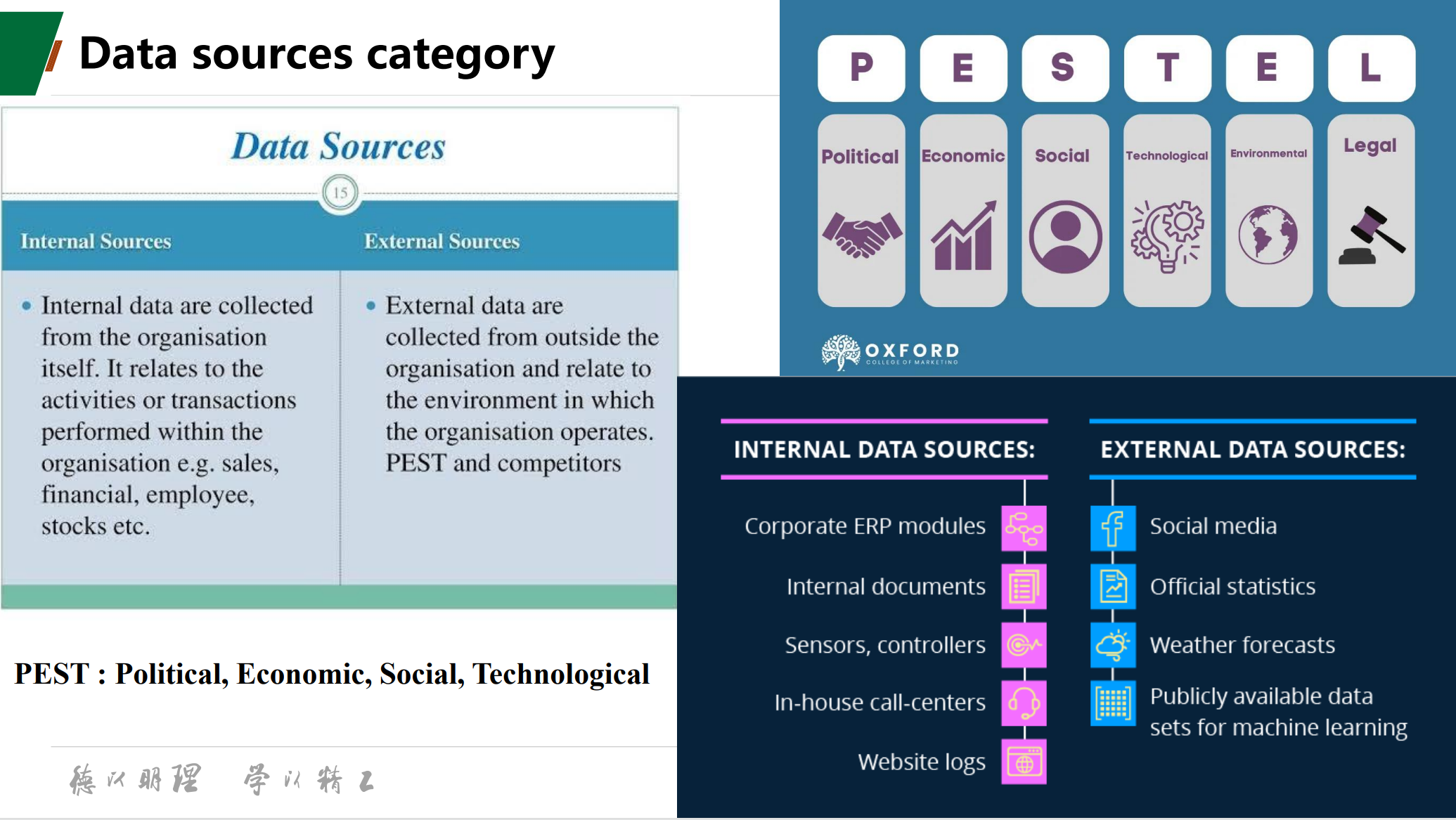
9. 持续优化与反馈

* 反馈机制：根据数据反馈调整和优化模型，确保持续改进。
* A/B测试：验证不同决策或方案的效果，进一步优化策略。

10. 行业应用案例

* 医疗：通过大数据分析健康数据，提供个性化治疗方案。
* 金融：根据数据预测市场动向、客户信用等，优化风险控制。
* 零售：通过分析顾客行为，优化产品推荐和库存管理。
* 智能制造：通过物联网数据监控设备，提升生产效率。

3 and 8 题（不放这张图差点忘了我是全英班的）



### 4. 如何实现企业内部数据价值？

要实现企业内部数据的价值，通常需要通过以下几个步骤：

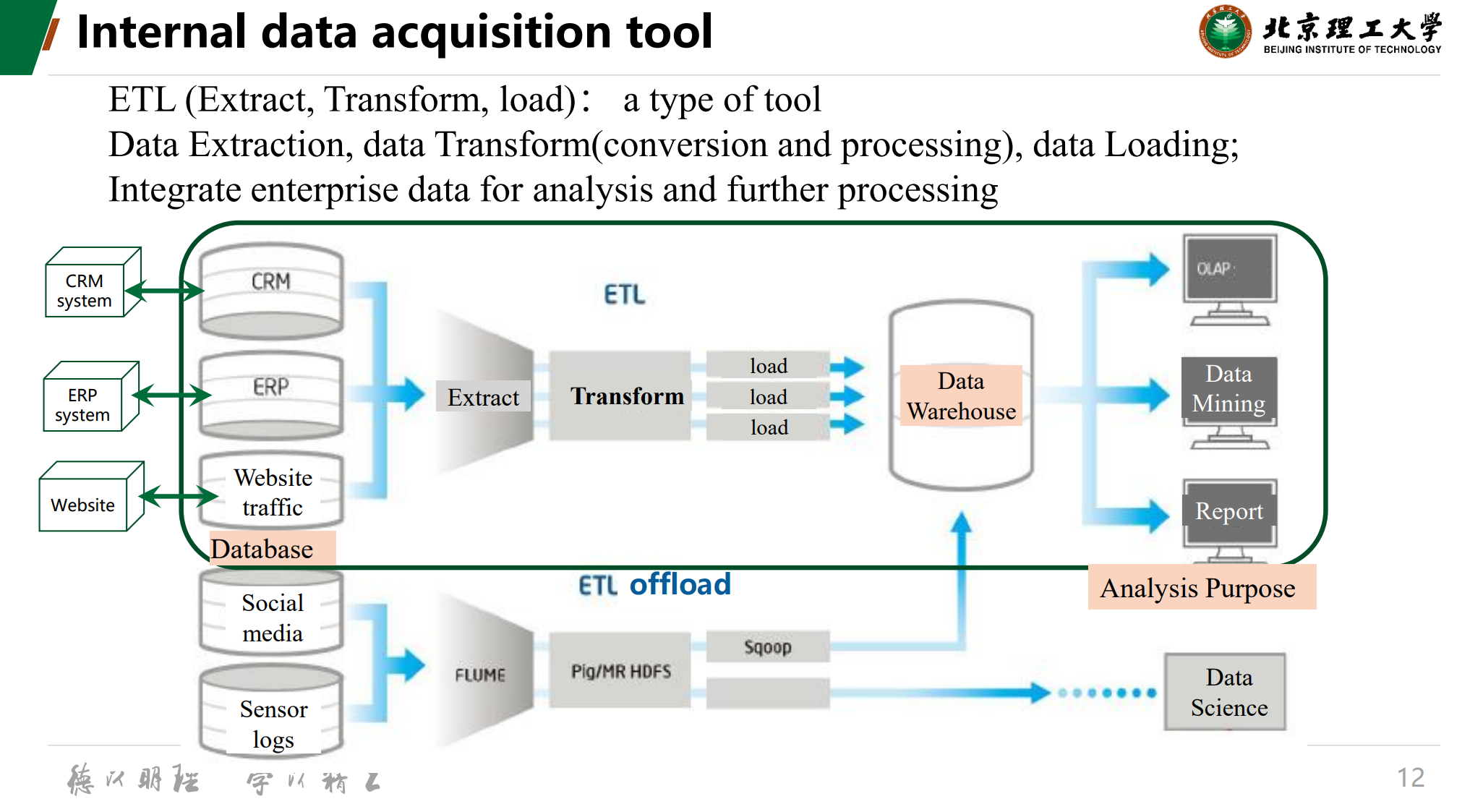
1. \*\*数据整合\*\*：将分散在各部门、系统和平台的数据进行整合，确保数据的统一性和一致性。

2. \*\*数据分析\*\*：运用数据分析方法，提取有价值的信息，发现潜在的商业机会、优化流程或改善决策。

3. \*\*数据可视化\*\*：通过可视化工具将复杂数据转化为易于理解的图表或报告，帮助决策者更快速地洞察业务状况。

4. \*\*数据驱动决策\*\*：基于分析结果，将数据转化为实际行动，推动战略决策和优化操作。

5. \*\*数据文化建设\*\*：培养员工的数据思维，使其能主动识别、分析和利用数据，以提升企业整体的敏捷性和创新能力。



### 5. 有哪些常用的组织内部数据获取方法？

常用的内部数据获取方法包括：

1. \*\*业务系统日志\*\*：通过企业的ERP、CRM、财务系统等，提取与业务活动相关的日志数据。

2. \*\*员工输入\*\*：通过调查问卷、反馈表、数据输入表格等方式获取员工提供的操作数据或意见反馈。

3. \*\*传感器数据\*\*：物联网设备、生产线传感器等实时采集数据，用于监控生产、环境等因素。

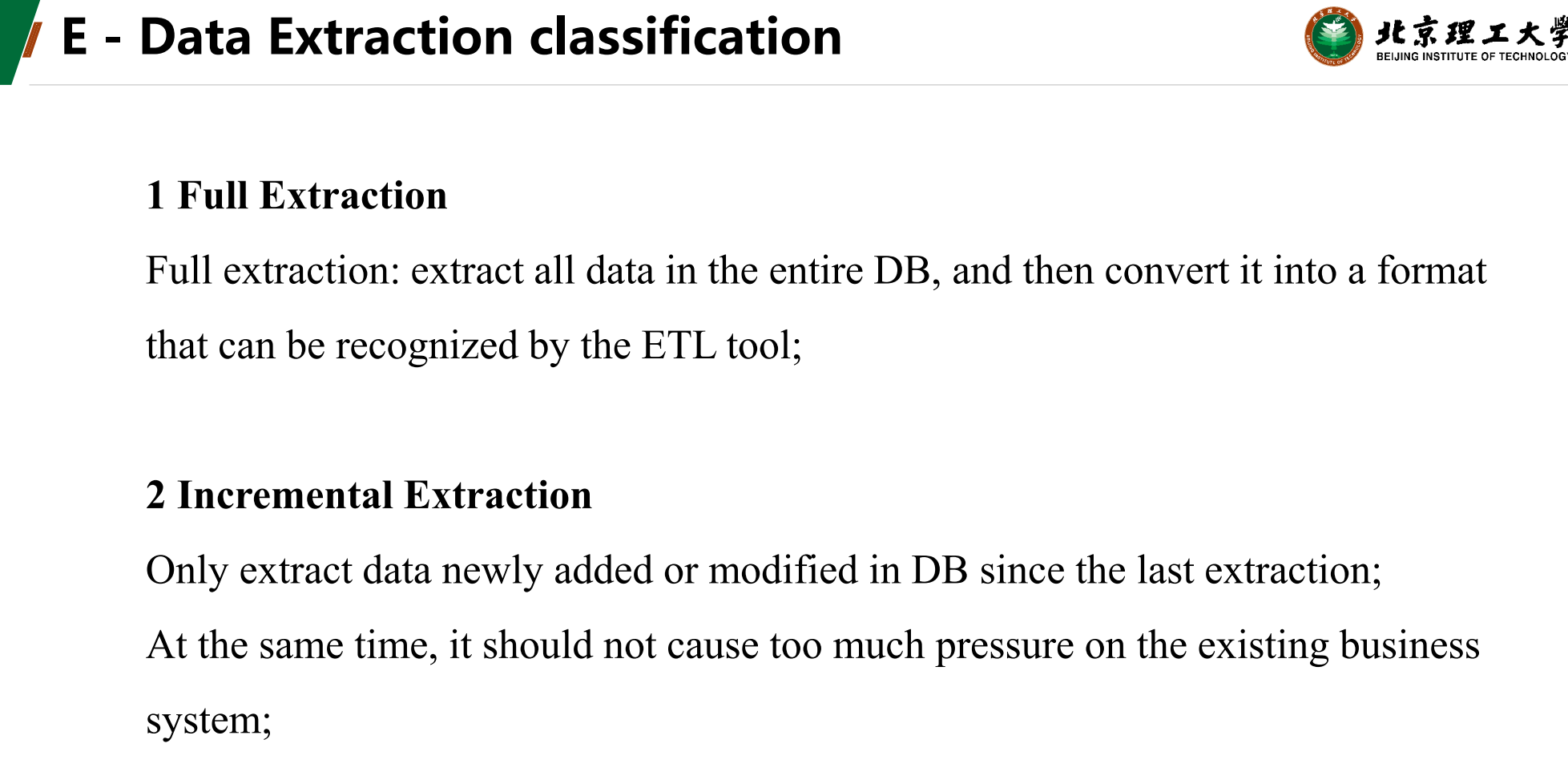
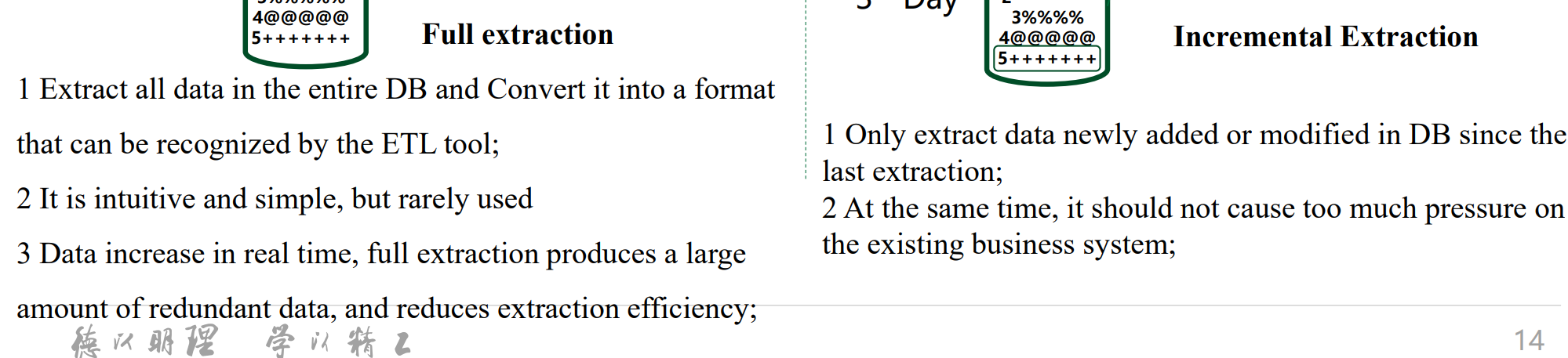
4. \*\*数据库查询\*\*：通过直接查询组织内部数据库，获取结构化数据。

5. \*\*文件和文档管理系统\*\*：通过文档管理系统提取存储的文档或表格数据。

### 6. 什么是数据抽取?有哪些常用的数据抽取方法？

\*\*数据抽取\*\*是从不同的数据源中提取原始数据的过程，以供后续的数据清洗、转换和分析。

常用的数据抽取方法包括：

1. \*\*数据库抽取\*\*：通过SQL查询语句从数据库中提取数据。

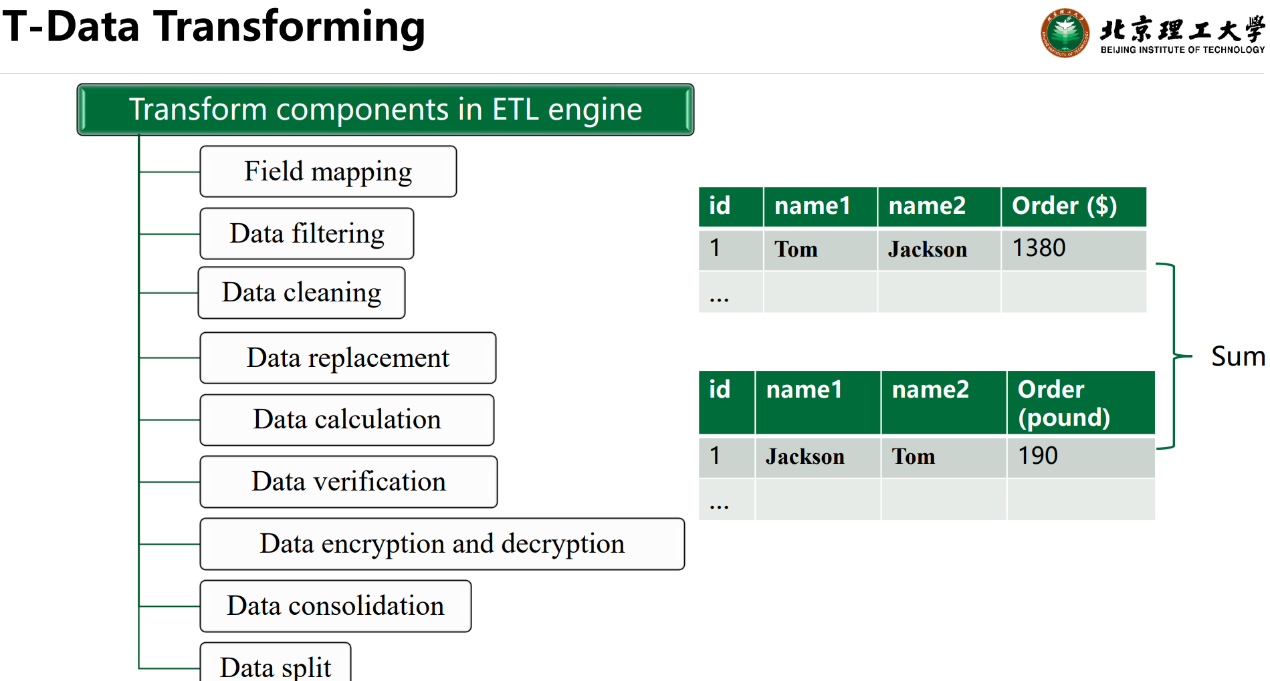
2. \*\*API调用\*\*：通过API接口从其他系统或第三方平台获取数据。

3. \*\*文件抽取\*\*：从Excel、CSV、JSON等文件格式中提取数据。

4. \*\*Web抓取\*\*：使用爬虫技术从网页中提取数据。

5. \*\*传感器数据采集\*\*：通过设备传感器实时获取数据，如温度、湿度等。

### 7. 什么是数据转换？



\*\*数据转换\*\*是指将抽取到的数据从一种格式转换为适合分析、存储和处理的另一种格式。数据转换的目标是确保数据的准确性、一致性，并适应目标系统的需求。

数据转换的常见操作包括：

1. \*\*数据清洗\*\*：删除或修正错误、缺失、重复的数据。

2. \*\*数据格式化\*\*：将数据转换成统一的格式，如日期格式、货币单位等。

3. \*\*数据标准化\*\*：将不同来源的数据统一为标准单位或标准类别。

4. \*\*数据集成\*\*：将来自不同源的数据整合到一个统一的数据集。

5. \*\*数据汇总\*\*：将数据汇总、聚合成更高层次的统计数据，如总和、平均值等。

### 8. 什么是组织外部数据?组织外部数据包含哪些内容？

见上文3 and 8

\*\*组织外部数据\*\*是指企业外部来源的数据，这些数据通常不属于企业内部生成或控制的数据，但对企业运营有重要影响。

外部数据的主要内容包括：

1. \*\*市场数据\*\*：包括行业趋势、竞争者信息、市场调研等。

2. \*\*社会经济数据\*\*：例如政府发布的经济指标、人口统计数据、消费者行为等。

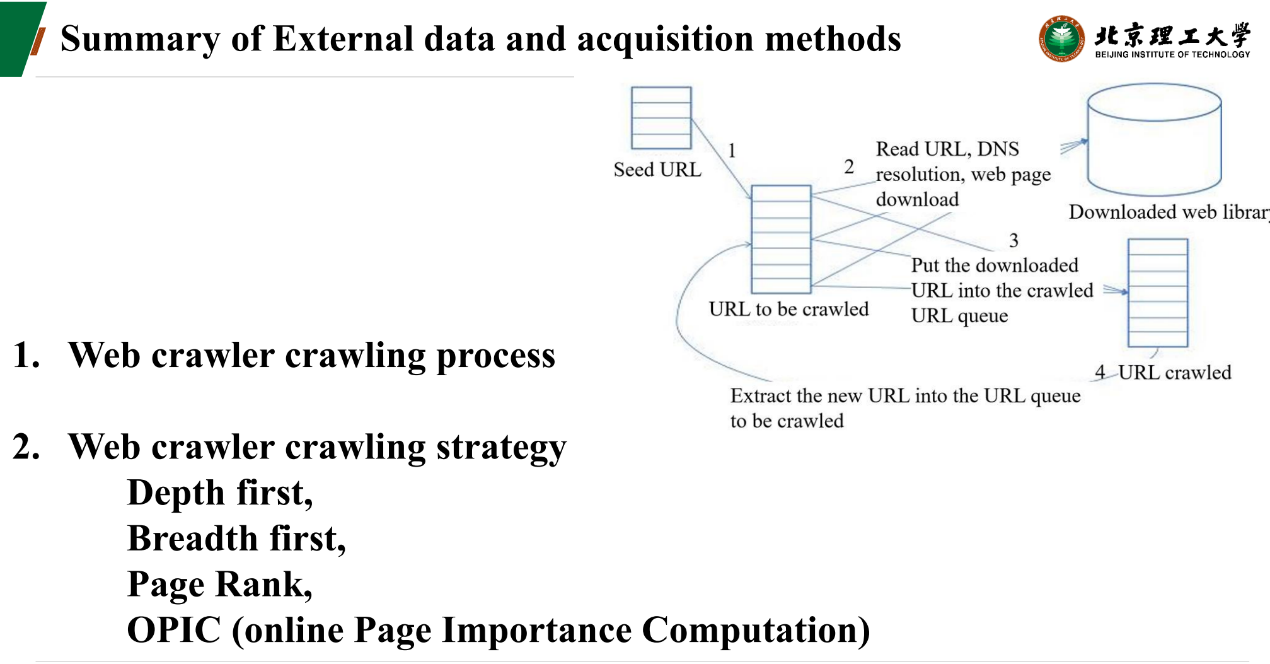
3. \*\*金融数据\*\*：如股票市场数据、金融报表、汇率、利率等。

4. \*\*地理位置数据\*\*：地图数据、交通状况、天气数据等。

5. \*\*社交媒体数据\*\*：从社交平台上获取的用户评论、情感分析等信息。

6. \*\*供应链数据\*\*：供应商和客户的相关信息。

### 9. 组织外部数据按照获取方法可以分为哪两类？



组织外部数据可以按照获取方法分为两类：

深网数据和浅网数据

### 10. 浅网数据获取方法是什么？

\*\*浅网数据\*\*是指互联网上公开可访问的、能够通过常规搜索引擎查找到的数据。获取方法包括：

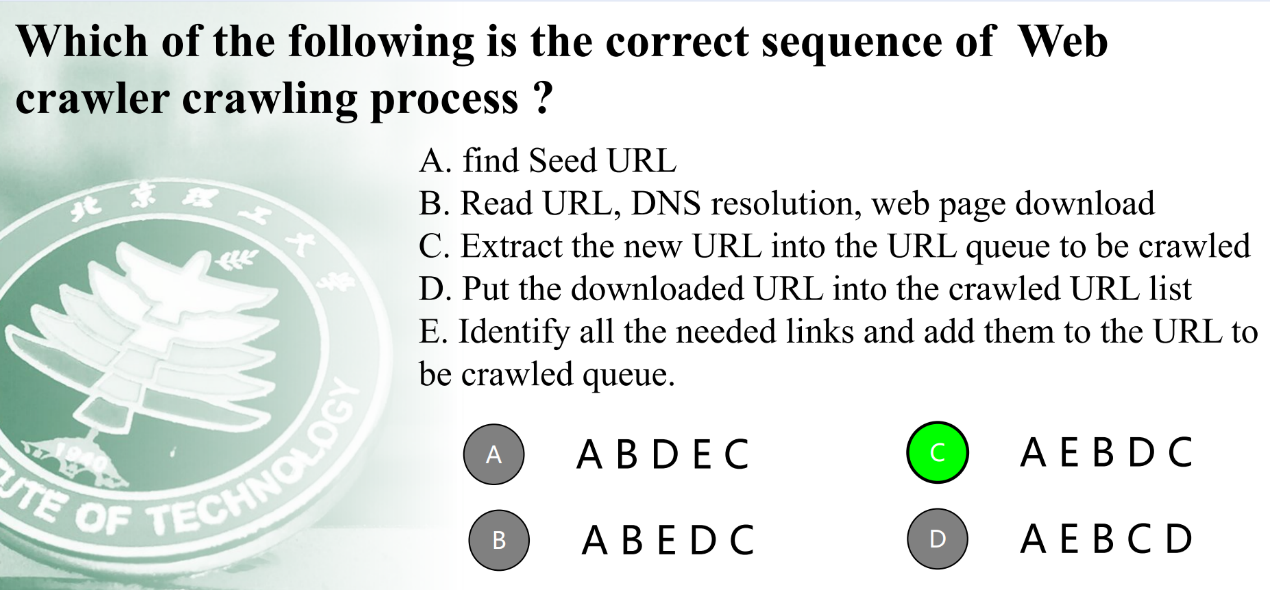
1. \*\*搜索引擎爬虫\*\*：使用Google、Bing等搜索引擎搜索特定信息。

2. \*\*公开API接口\*\*：如社交平台的API，提供公开的业务数据、趋势信息等。

3. \*\*手动收集\*\*：通过访问公开网页、报告、新闻稿等方式获取数据。

4. \*\*开放数据集\*\*：许多政府或机构提供的公开数据集，可以直接下载使用。

### 11. 网络爬虫爬取流程是什么？



网络爬虫的爬取流程通常包括以下步骤：

1. \*\*爬虫初始化\*\*：设置爬虫的起始URL、目标网站和抓取规则。

2. \*\*发送HTTP请求\*\*：爬虫向目标网站发送HTTP请求，获取网页内容。

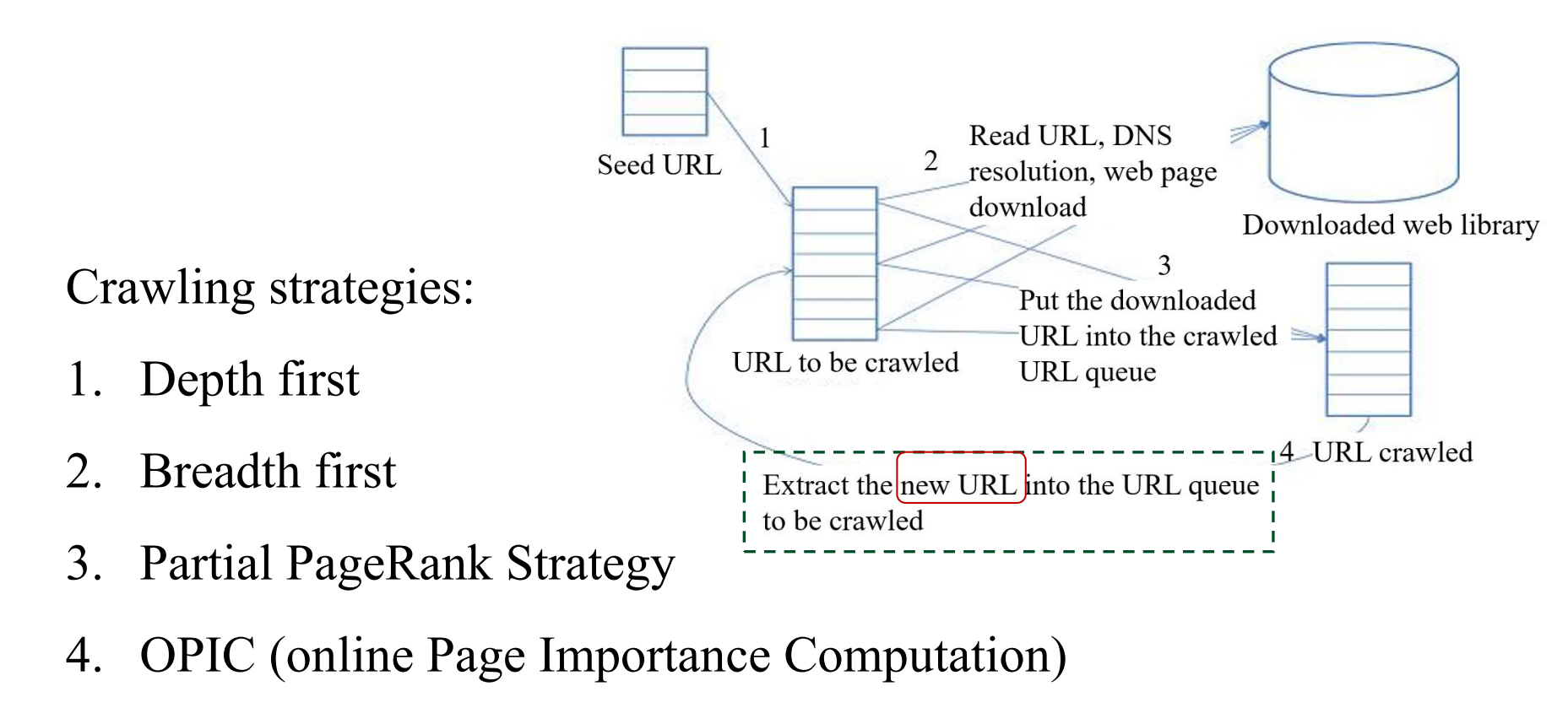
3. \*\*网页解析\*\*：解析网页的HTML结构，提取所需的数据（如文本、图片、链接等）。

4. \*\*数据存储\*\*：将提取的数据保存到数据库或文件中。

5. \*\*URL调度\*\*：识别网页中的其他链接并将其添加到待爬取的URL列表中，进行下一轮抓取。

6. \*\*反爬虫策略处理\*\*：绕过网站的反爬虫机制，如IP封锁、验证码等。

### 12. 有哪些常见的网络爬虫策略？



常见的网络爬虫策略包括：

1. \*\*深度优先搜索（DFS）\*\*：从当前页面开始，一直深入到页面的最底层，再回溯爬取其他页面。

2. \*\*广度优先搜索（BFS）\*\*：从当前页面开始，首先抓取所有直接链接的页面，然后再逐步深入。

3. \*\*延时抓取\*\*：为了避免被目标网站的反爬虫系统封锁，可以设置抓取延时。

4. \*\*分布式爬虫\*\*：通过多个爬虫节点同时抓取，提高抓取效率。

5. \*\*IP轮换\*\*：使用代理服务器轮换IP地址，防止被网站封锁。

6. \*\*User-Agent伪装\*\*：改变请求头中的User-Agent信息，模拟不同的浏览器行为。

### 13. 什么是深网数据?如何获取深网数据？

\*\*深网数据\*\*是指那些未被常规搜索引擎索引的数据，通常存储在数据库、私密网站、动态网页等非公开页面上。深网数据包括但不限于：

- 数据库记录

- 动态生成的网页（例如通过JavaScript生成内容的网页）

- 受密码保护的页面

- 内部系统数据等

\*\*获取深网数据的方法：\*\*

1. \*\*直接访问数据库\*\*：通过合法权限直接访问深网数据库。

2. \*\*API接口\*\*：使用API接口获取深网中的数据，通常需要注册并认证。

3. \*\*Web Scraping（动态网页抓取）\*\*：使用爬虫模拟浏览器，抓取JavaScript动态加载的数据。

4. \*\*VPN与代理\*\*：绕过地域或IP限制，访问深网内容。

深网与\*\*暗网\*\*（Dark Web）不同，后者通常涉及非法或高度隐秘的数据和活动。

习题三

3-1数据管理系统的发展经历了哪些阶段?

3-2大数据管理系统具有哪些典型的系统特征?

3-3大数据管理系统中有哪些常见的大数据模型?

3-4什么是键值对模型?哪些应用适合采用键值对模型进行数据存储?

3-5 什么是列族模型?哪些应用适合采用列族模型进行数据存储?

3-6 什么是文档模型?哪些应用适合采用文档模型进行数据存储?

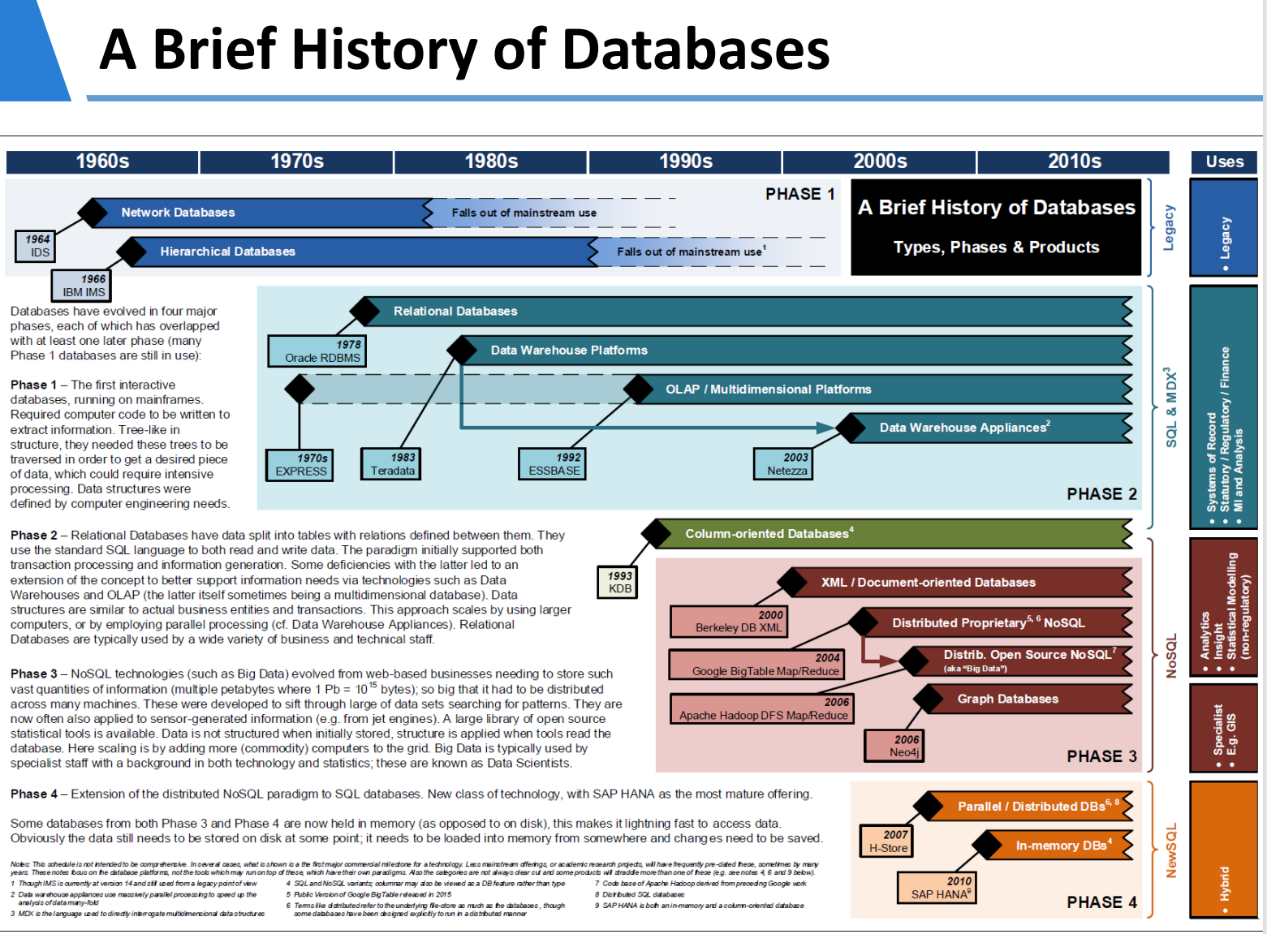
3-7什么是图模型?哪些应用适合采用图模型进行数据存储?

3-8分布式文件系统HDFS如何进行大数据组织和管理?

3-9 相比于关系数据库，NoSQL数据库具有哪些优势?

3-10 何为NewSQL数据库?

（这一部分的习题看起来正常的多）

1. 

2. 

3-7全部见下图



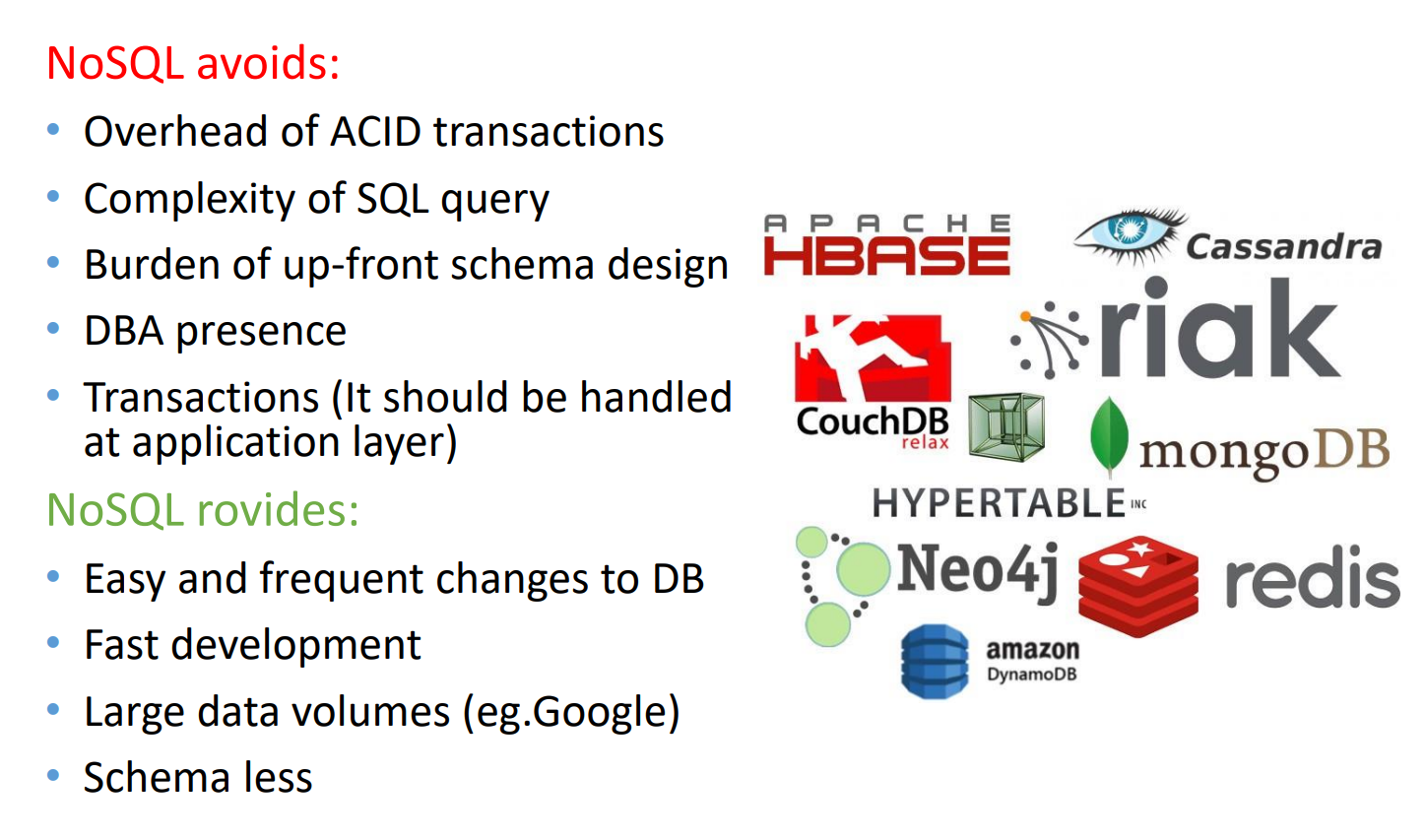
8.• In traditional approach, all the data were stored in a single central machine. With the rise of big data, a single database is not enough for storage. • The solution is to use a distributed approach to store the massive amount of data. Data are divided and distributed among many nodes (machines).

Distribution: Distribute blocks of data sets across multiple nodes.

Replication: Replicate data blocks on different nodes by copying the same pieces of information into multiple nodes on different racks. • Fault Tolerance: recover data block in case of cluster/rack failure • High Concurrency: same data can be processed by multiple clients at the same time.

HDFS splits massive files into small chunks, these chunks are known as data blocks • All the blocks are of the same size but the last block can be of the same size or lesser

HDFS blocks are replicated multiple times • One block stored at multiple location, also at different racks (usually 3 times) • This makes HDFS storage fault tolerant and faster to read

9. A No SQL database provides a mechanism for storage and retrieval of data that employs less constrained consistency models than traditional relational database ❑No SQL systems are also referred to as "NotonlySQL“ to emphasize that they do in fact allow SQL-like query languages to be used.

10 NewSQL 是对各种新的可扩展/高性能数据库的简称，这类数据库不仅具有[NoSQL](https://baike.baidu.com/item/NoSQL/0?fromModule=lemma_inlink" \t "_blank)对海量数据的存储管理能力，还保持了传统[数据库](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BA%93/0?fromModule=lemma_inlink" \t "_blank)支持[ACID](https://baike.baidu.com/item/ACID/0?fromModule=lemma_inlink)和[SQL](https://baike.baidu.com/item/SQL/0?fromModule=lemma_inlink" \t "_blank)等特性。

NewSQL是指这样一类新式的关系型数据库管理系统，针对OLTP（读-写）工作负载，追求提供和NoSQL系统相同的扩展性能，且仍然保持ACID和SQL等特性（scalable and ACID and (relational and/or sql -access)）。

习题四

4-1请简述大数据处理生态系统的层次化结构中包含的主要功能组件，

4-2请列举典型的大数据处理模式。

4-3请简述批处理模型的概念，并列举典型的代表框架

4-4请举例说明适合采用批处理模型的应用。

4-5请简述流处理模型的概念，并列举典型的代表框架

4-6请举例说明适合采用流处理模型的应用。

4-7请简述图计算模型的概念，并列举典型的代表框架。

4-8请举例说明适合采用图计算模型的应用。

4-9请简述你对 MapReduce计算框架的理解(无须描述具体执行过程)

4-10请简述Spark系统相比于Hadoop系统存在的优势

习题五

5-2简述大数据分析常用方法。

5-2什么是联邦学习?

5-3简述分布式学习中通信步调的类型。

5-4数据可视化有什么作用?

5-5简述分布式系统的组成结构。

5-6简述基于数据流图计算模型的大数据分析框架的种类

5-7简述采用数据流图计算的优势。

5-8简述 ps-list参数服务器的分布式流程

5-9简述批流混合处理的意义。

5-10 简要描述TensorFlow的计算图。

习题六

6-1 如何理解大数据是一把双刃剑?

6-2数据安全解决方案必须满足的3个要求是什么?

6-3 大数据安全保护主要涉及大数据使用的哪些环节?

6-4 大数据平台安全的组成有哪些?各自有什么作用?

6-5 隐私保护的技术手段有哪些?

6-6目前有哪些技术可以实现安全多方计算?这些技术之间有哪些联系与区别

6-7 联邦学习中其他设备是否可以访问自己的本地数据?

6-8 国际主流的数据安全治理框架有哪些?它们分别有什么特点?

6-9大数据防护的策略有哪些?有何重要意义?

习题七

7-1 请简述大数据治理的概念。

7-2 请简述大数据治理中涉及的参与主体

7-3 请简述大数据治理的核心技术。

7-4 请简述元数据的概念

7-5 请简述元数据及元数据管理的重要性

7-6 请简述主数据管理的重要性

7-7请简述数据质量的几个重要维度

7-8 请简述数据质量管理常见的技术手段

7-9 请简述数据集成常用的技术手段

7-10 请举例说明数据标准化的应用。