

大数据导论 考察

一、简答题

1. 简述什么是结构化数据、非结构化数据、半结构化数据

结构化数据是指可以使用关系型数据库表示和存储，表现为二维形式的数据。结构化数据遵循一个标准的模型或者模式，并且常以表格的形式存储；

非结构化数据是数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据，即不遵循统一的数据模式或者模型的数据；

半结构化数据有一定的结构与一致性约束，但本质上不具有关系性。

2. 简述大数据的五大特征

大数据的五大特征指5V特征，由IBM提出，即：Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）。

3. 简述什么是数据集、数据分析、数据分析学

数据集就是数据的集合，所谓的集合表现就是将数据以某种方式统一存储在一定形式的数据库中；

数据分析是有组织有目的地收集数据、分析数据，使之成为信息的过程；

数据分析学是有关数据分析、研究并改善数据分析的一门学科。

4. 简述大数据分析的生命周期可分为哪些阶段

大数据分析的生命周期分为六个阶段：

1. 发现

2. 数据准备（分析沙箱）

3. 模型规划（方法，技术，工作流程，变量，关系，模型）

4. 建立模型（训练和测试数据集，软硬件）

5. 沟通结果（确定关键发现）

6. 实施（交付，试点项目）

5. 简述大数据与云计算、物联网、人工智能的关系

物联网、大数据、人工智能、云计算，作为当今信息化的四大版块，它们之间有着本质的联系，具有融合的特质和趋势。物联网是基础中的基础，为各类服务提供硬件上的基本支持，大数据是基于物联网的应用，也是人工智能的基础，而人工智能又是大数据的最理想应用，还能改善大数据，而云计算为各种服务提供统一的平台，是一切的依托。

6. 简述什么是数据安全和隐私保护

数据安全是指就是承载信息的数据的安全，在大数据应用中，则是指用各种手段保护其所涉及的结构化数据和非结构化数据等的安全。

由于在进行大数据分析时，可能会透露出一些组织或个人的机密信息，因此可能导致一些有意或无意的隐私暴露。因此在进行大数据应用时，需充分保证数据的安全，在此基础上，隐私保护也基本上能做到。

隐私保护是合规驱动，数据安全是业务驱动，但是随着大家对数据安全和隐私保护的重视，和国家的不断监管指引，数据安全和隐私保护，一定会变成业务主动驱动，成为企业的核心竞争力。

二、论述题

1. 试述定量分析、定性分析、数据挖掘、统计分析、机器学习、语义分析、视觉分析等数据分析技术的基本概念

1. 定量分析：是一种数据分析技术，专注于量化从数据中发现的模式和关联，样本容量要求大，结果可在整个数据集中推广。

2. 定性分析：专注于用语言描述不同数据的质量，涉及样本相对较小且深入，其结果不适用于整个数据集，也不是和测量数值或用来比较。

数据挖掘：也叫数据发现，是一种针对大型数据集的数据分析的特殊形式，一般是自动的、基于软件技术的、筛选海量数据集来识别模式和趋势的技术。

3. 统计分析：即以数学公式为手段来分析数据，统计方法可以是定量的，也可以是定性的，可以用于推断数据集中的模式和关系。

4. 机器学习：是指从有限的观测数据中学习 (或“猜测”) 出具有一般性的规律，并将这些规律应用到未观测数据样本上的方法。

5. 语义分析：是自然语言处理的一种实现方式，尝试通过各种方法分析多种媒介承载的词语、句子或篇章的内在含义，提取有意义的信息。

6. 视觉分析：指对数据进行图形表示来开启或增强视觉感知，目标是用图形表示来开发对分析数据的更深入的理解。

2. 学习打数据导论课程后，有何体会（答题建议：1. 根据各国的大数据发展战略，你如何看待大数据的发展前景；2. 畅想大数据在未来还会有哪些重要的应用场景；3. 畅谈未来你希望涉足的大数据具体应用领域以及计划）

Mine：基于日常语义分析、图像识别、音频分析和应用列表获取对潜在犯罪人员的自动化监控系统