**平 时 作 业 报 告**

**课程名称： 大数据与云计算导论**

**学院： 计算机与软件学院**

**专业： 软件工程**

**指导教师： 陆玉武**

**报告人： 郑彦薇 学号： 2020151022**

**时间： 2021年3月22日**

**报告提交时间： 2021年3月24日**

|  |
| --- |
| **说明：**  （1）本次作业满分为100分。  （2）本次作业截至时间2021年3月30日（周二）23:59。  （3）个人信息：WORD文件名中的“姓名”、“学号”，请改为你的姓名和学号；报告的首页，请准确填写“学院”、“专业”、“报告人”、“学号”、“实验报告提交时间”等信息。  （5）提交方式**：截至时间前，请在**Blackboard系统中提交，延迟提交无效。  （6）发现抄袭（包括复制&粘贴整句话、整张图），抄袭者和被抄袭者的总成绩记零分，可以参考网络资源。  （7）期末考试阶段补交无效。  （8）提交作业的命名方式为：**大数据与云计算导论-学号-姓名**，请按规定命名并按时提交作业。 |
| **（1）请举例说明结构化数据、半结构化数据、非结构化数据的区别。（10分）**  答：**（1）结构化数据是指具有较强的结构模式，可以使用关系型数据库表示和存储的数据；（2）半结构化数据是一种弱化的结构化数据类型；（3）非结构化数据不具固定的数据结构，各种存储在文本文件中的系统日志、文档、图像、音频、视频等数据都属于非结构化数据。**  【例子】：结构化数据：    半结构化数据：    非结构化数据：图片、音频等。  **（2）什么是大数据的4V或5V特征？这一特征对大数据计算过程带来什么样的挑战？（15分）**  答：**4V特征，即规模庞大，种类繁多，变化频繁和价值巨大但价值密度低。**  **5V特征，在4V特征的基础上增加了其他提法，强调数量，多样性，速度和真实性等方面后将数据价值吸引进来，形成大数据的所谓“5V特征”。**  **挑战：对发现事实、揭示规律并预测未来提出了新的挑战，并将对已有计算模式、理论和方法产生深远的影响。**  **（3）请在生活中举出一个基于“数字映像”探索或研究现实世界的实体或现象的例子，思考如何才能获得更准确的数字映像？（15分）**  答：**区别于传统统计学，人们有可能通过获得更接近于全样的抽样，形成一个客观世界的实体和现象在计算机能够处理信息世界中的一个数字映像。**  **例子：在智能制造系统中，有数字孪生的概念，美国国防部最早提出在数字空间里建立真实飞机的模型，并通过传感器实现与飞机真实状态的完全同步。这样，每次飞行后，就可以基于数字模型的现有情况和过往载荷，及时分析评估飞机是否需要维修，能否承载下次任务载荷等。**  **如何才能获得更准确的数字映像：将数据储存下来，拥有真实数据以及对数据的实时处理能力，从大量无序的数据中获取价值**   1. **大数据集成的基本原理有哪些？数据质量有几种维度？分别是什么？（15分）**   答：（1）**大数据集成的基本原理：大数据集成应用，首先是从企业内部的小数据开始起步的；**  **首先搭建企业内部的大数据平台，可以用开源框架，可以用Ambari集成工具实践；**  **基于本企业的业务数据建立Key指标，首先实现这些指标的准实时统计和可视化；**  **基于内外部的更多数据收集，建立企业需要的更高级的数据模型，从横向和纵向进行拓展和深挖，达到指导企业战术和战略的预测与指导作用。**   1. **数据质量的维度：有5种维度，分别是：**   **数据一致性：在数据集合中，每个数据都不包含语义错误或相互矛盾的数据；**  **数据完整性：数据集合中包含足够的数据来回答各种查询，并支持各种计算；**  **数据精确性：数据集合中，每个数据都能准确表述现实世界中的实体；**  **数据时效性：在信息集合中，每个信息都与时俱进，保证不过时；**  **实体同一性：同一实体的标识在所有数据集合中必须相同而且数据必须一致。**  **（5）大数据的来源主要有几种？如何理解数据科学。（15分）**  答：**大数据的来源：对现实世界的测量，人类的记录和计算机生成这三种；**  **数据科学是将数据转化为决策和行动的艺术，是人和计算机一起工作将数据转化为知识发现的工具、技术和流程的整合。主要有两个内涵：一个是研究数据本身，研究数据的各种类型、状态、属性及变化形式和变化规律；一个是为自然科学和社会科学研究提供一种新的方法，称为科学研究的数据方法，其目的在于揭示自然界和人类行为现象和规律。**  **（6）大数据集成的基本原理有哪些？你能提出一个教育行业领域中的数据获取的应用案例吗？（10分）**  **答：**（1）**大数据集成的基本原理：大数据集成应用，首先是从企业内部的小数据开始起步的；**  **首先搭建企业内部的大数据平台，可以用开源框架，可以用Ambari集成工具实践；**  **基于本企业的业务数据建立Key指标，首先实现这些指标的准实时统计和可视化；**  **基于内外部的更多数据收集，建立企业需要的更高级的数据模型，从横向和纵向进行拓展和深挖，达到指导企业战术和战略的预测与指导作用。**  **（2）案例：南京理工大学利用大数据资助贫困学生：南京大学教育基金会通过数据分析，每个月在食堂吃饭超过60顿、一个月消费不足420元的，被列为受资助对象。他们采用直接将补贴款打入学生饭卡的形式，学生无需填表申请。**  **（7）请分析在交通大数据(如GPS采集的数据、打车软件中记录的数据)中可能遇到数据质量问题以及这些数据质量问题的检测方法和修复方法。（20分）**  **答：鉴于数据质量的重要性，数据质量管理已经成为一个学科。数据质量问题可以通过制度手段和技术手段解决。制度手段包括指定数据质量度量标准、数据质量监管体系和数据质量管理制度等，技术手段包括一系列错误检测和修复方法。**  **主要有：**  **不一致的检测和修复：基于数据完整性约束，给定一组完整性约束，对于数据库中不满足约束的部分I，通过修复操作求解与该部分差距最小的I’,且I’满足约束。**  **缺失值填充：针对不完整数据进行缺失数据的填充，有删除、统计填充、统一填充、预测填充这几种常用的方法。**  **实体识别和真值发现：实体识别是指在给定的实体对象（包括实体名和各项属性）集合中，正确发现不同的实体对象，并将其聚类，使得每个经过实体识别后得到的对象簇在现实世界指代的是同一个实体。**  **真值发现是指不同数据源对同一实体可能提供不同的数据冲突，这就需要在冲突数据中找出真值。** |
|  |

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |