**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5月12日**

**报告提交时间： 2021/5/17**

|  |
| --- |
| **说明：**  （1）本次作业满分为100分。  （2）本次作业截至时间2021年5月26日（周三）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）给出可视化的定义，并说明结构化数据的类型，非结构化数据的类型。（10分）**  答：**可视化的定义：**可视化是理解、探索、分析大数据的重要手段。它通过将数转化为图形图像提供交互，以帮助用户更有效地完成数据的分析、理解等任务。可视化可以迅速有效地简化与提炼大量的数据，并从中寻找新的线索，发现和创造新的理论、技术和方法，改善大众的日常生活。  **结构化数据的类型：**结构化数据包括四个基本类型：表格数据、网络数据、场数据和几何数据。  **非结构化数据：**非结构化数据包括自然语言文本、图片、视频等，常需要转换为结构化的数据以便进行可视化。  **（2）给出一组一维数据均值、中位数、及标准差的定义。（10分）**  答：**均值的定义**：一组数据共有n个一维数据，均值为这些数据的总和除以个数。  **中位数的定义：**中位数是数据中按照大小顺序排序后，处于中间位置的数；一共有n个数，若n为奇数，则中位数为位于最中间的那个数，若n为偶数，则中位数为位于中间两个数的平均数。  **标准差的定义：**方差的算术平方根就是标准差s。方差是数据中每一个数与平均数作差求平方后求和，再除以总的数据的个数得到的结果。  **（3）什么是数据加密技术？PKI的体系主要结构有哪些？（15分）**  答：**数据加密技术：**数据加密是一个利用加密密钥通过加密算法将明文信息转换成密文信息的处理过程，收到密文的接受方利用解密密钥通过解密算法还原成明文。  **PKI的体系主要结构：**认证机构、证书和证书库管理、密钥备份及恢复管理、多个PKI间的交叉认证、时间戳、不可抵赖机制。  **（4）什么是流式大数据计算？流式大数据计算与批量大数据计算的主要区别有哪些？**  **（15分）**  答：**流式大数据计算的定义：**流式计算主要面向在线计算场景，计算的数据是动态数据，数据在计算过程中不断的到来，计算前无法预知数据的到来时刻和到来顺序，也无法预先将数据进行存储。通常流式计算是数据边到来边计算，计算的实时性要求高。流式大数据计算系统通常是一个数据处理拓扑或管道。该系统包括数据源节点、数据处理节点和数据分发逻辑。  **流式大数据计算与批量大数据计算的主要区别：**  相对于批量计算，流式计算具有如下特点：**实时性：**流式大数据是实时产生、实时计算，结果反馈往往也需要保证及时性；**易失性：**在流式大数据计算环境中，数据流往往是到达后立即被计算并使用，只有极少数的数据才会被持久地保存下来，大多数数据往往会被直接丢弃；**突发性：**在流式大数据计算环境中，数据地产生完全由数据源确定，由于不同的数据源在不同时空范围内的状态不统一且发生动态变化，导致数据流的速率呈现处了突发性的特征；**无序性：**在流式大数据计算环境中，各数据流之间、同一数据流内部各数据元素之间是无序的；**无限性：**在流式大数据计算中，数据是实时产生的、动态增加的，只要数据源处于活动状态，数据就会一直产生和持续增加下去。  **（5）图数据计算有哪些典型特征 ？简述BSP模型的基本原理和BSP计算的主要步骤。**  **（15分）**  答：**图数据计算的典型特征：**图数据是一种重要而且普遍的大数据，它是研究物体与物体之间的关系，并进行整体的刻画、计算和分析的一种技术。图计算存在较强的局部依赖性，使得图计算具有局部更新和迭代计算的特征。  **BSP模型的基本原理：**BSP是一种异步多指令流多数据流-分布式存储模型，一个BSP并行计算系统由一组通过通信网络互连的处理器与内存单元组成。主要有三个部分：一组具有局部内存的分布式处理器、全局数据通信网络、支持所有处理单元间全局栅栏同步的机制。  **BSP计算的主要步骤：**一次BSP计算过程由一系列全局超步组成。超步就是BSP计算中的一次迭代。局部并发计算：由每个参与计算的处理器执行计算任务，负责读取和处理存储在本地内存中的数据；通信：处理器之间相互交换数据，由一方发起推送或获取操作来实现数据交互；栅栏同步：当处理器遇到栅栏时，会等其他所有处理器完成他们的计算步骤；每一次同步是一个超步的完成，也是下一个超步的开始。  **（6）工业大数据包含工业企业的哪些数据源？工业大数据的特点是什么？工业大数据的分析流程是怎样的？（15分）**  答：**工业大数据包含的数据源：**工业大数据是工业企业自身及生态系统产生或使用的数据总称，既包含企业内部来自CAX、MES、ERP等信息化系统的数据、生产设备、智能产品等物联网数据，也包括企业外部来自上下游产业链、互联网以及气象、环境、地理信息等物联网数据，贯穿于研发设计、生产制造、售后服务、企业管理等各环节。  **工业大数据的特点：本质特征：**工业系统往往具有复杂动态系统特性；确定性是工业系统本身能够有效运行的基础；工业系统同时具有确定性和不确定性的特征，确定性是目标，不确定性是机会。**5V特征：**工业大数据复合大数据的5V特征，即大规模、速度快、类型杂、低质量、低价值。**新特征：**工业大数据作为对工业相关要素的数字化描述和在赛博空间的映像，除了具有大数据的5V特征，相对于其他类型的大数据，工业大数据集还具有反映工业逻辑的新特征，这些特征可以归纳为多模态、强关联、高通量。**应用特征：**工业大数据的应用特征可归纳为跨尺度、协同性、多因素、因果性、强机理等几个方面，这些特征都是工业对象本身的特性或需求所决定的。  **工业大数据的分析流程：**工业大数据的分析工作可大致分为分析工作准备和分析工作实施两个阶段。**分析工作准备**包括业务理解和数据理解两个步骤；**分析工作实施**包括数据准备、分析模型设计、验证与评估和实施与运营等几个步骤；整个工业大数据分析工作是一个逐步循环迭代的过程。  **（7）教育大数据的应用场景有哪些？除此之外，请再提供至少两个教育大数据的应用例子。（20分）**  答：**应用场景：**教育大数据的采集和应用场景涵盖了校园管理、教学辅导、学习资源、开放性论坛等教育环节中的方方面面；一般来说，教育场景包括信息化校园、智能辅导系统和在线题库、大规模开放式在线课程等主要类别；  **应用例子：（一）：**学校管理者可以管理学校的资产信息、师资信息、招生就业信息等。同时随着移动互联网以及物联网等新技术的兴起，学校师生主动产生和由设备自动收集的信息越来越多；  **（二）：**通过大数据帮助学生评估他们已有的知识和达标测验所需程度的差距，进而指出学生有待提高的地方，同时也可以让老师跟踪学生学习情况，从而找到学生的学习特点和方法。 |

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