试卷代号:1480　　座位号□□

国家开放大学2020年春季学期期末统一考试

大数据技术导论　试题

2020年9月

题　号	一	二	三	四	总　分
分　数					

得　分	评卷人

一、选择题(选择一个正确答案的代码填入括号中,每小题3分,共30分)

1. 由于数据随着时间而变化,可以将(　　)变化可视化,然后解释导致数据变化的原因。

A. 环境　　B. 时间

C. 数据　　D. 知识

2. 建立挖掘模型、选取或改进挖掘模型都需要验证,最常用的验证方法是(　　)。

A. 样本学习　　B. 统计分析

C. 逻辑推理　　D. 数学期望

3. 在样本数据较大的情况下,随机性越(　　),效果越好。

A. 弱　　B. 小

C. 高　　D. 低

4. (　　)是样本相对于均值的偏差平方和的平均。

A. 极差　　B. 变异系数

C. 标准差　　D. 样本方差

5. 数据约简主要有特征约简、样本约简、(　　)和数值约简等。

A. 维数约简　　B. 归一化

C. 数据变换　　D. 一致性

6. 大数据抽取过程就是从数据源中抽取数据并传送到(　　)中的过程。

A. 数据源　　B. 信息

C. 数据库　　D. 目的数据系统

7. NewSQL 适用于(　　)。

A. 事务处理应用　　B. 日志数据存储

C. 数据分析应用　　D. 互联网应用

8. 去重是指在不同的时间维度内,重复一个行为产生的数据只计入一次。按(　　)维度去重主要分为按小时去重、按日去重、按周去重、按月去重或按自选时间段去重。

A. 高维　　B. 低维

C. 时间　　D. 空间

9. 数据平滑法主要分为(　　)、指数平滑法和分箱平滑法。

A. 统计法　　B. 最短距离法

C. 移动平均法　　D. 聚类方法

10. 大数据的 5 个"V"特性是数据量、多样性、(　　)、速度、真实性。

A. 稀疏性　　B. 关联性

C. 实用性　　D. 价值

得　分	评卷人

二、判断题(正确的划√,错误的划×,每小题 2 分,共 20 分)

11. 同构同质数据库是指同一类型的数据模型、同一型号的数据库系统;同构异质数据库是指同一类型的数据模型、不同型号的数据库系统。(　　)

12. 数据规范化可将原来的度量值转换为无量纲的值,通过将属性数据按比例缩放,将一个函数给定属性的整个值域映射到一个新的值域中,即每个旧的值都被一个新的值替代。(　　)

13. 数据挖掘主要注重解决分类、聚类、关联和定量定性预测等问题,其重点不是寻找未知的模式与规律。(　　)

14. 一幅图画最伟大的价值莫过于它能够使我们实际看到的内容比期望看到的内容丰富得多。(　　)

15. 通过将抽象的指标数据转换成我们熟悉的容易感知的数据时,用户便更不容易理解图形要表达的意义。(　　)

16. 全量抽取类似于数据迁移或数据复制,它将抽取数据源中发生改变的地方数据从数据库中抽取出来,并转换成抽取工具可以识别的格式。(　　)

17. 只有通过清洗之后,才能通过分析与挖掘得到可信的、可用于支撑决策的信息。(　　)

18. 如数据不完整、数据不一致、数据重复等,数据也能够有效地被利用。(　　)

19. 采用 NoSQL+NewSQL 混合模式构建数据中心,可以发挥 NoSQL 数据库的事务处理能力和 NewSQL 在实时性、复杂分析、即席查询等方面的优势,以及面对海量数据时较强的扩展能力。(　　)

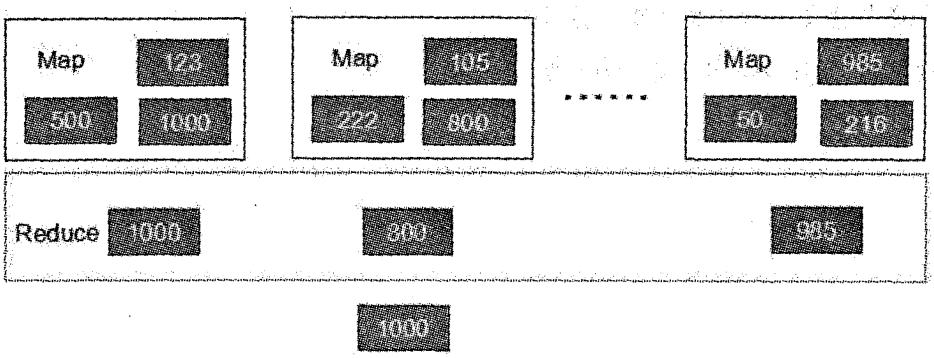
20. 数据科学的组成要素主要包括数学、统计学知识以及领域的专业知识。(　　)

得　分	评卷人

三、简答题(每小题 5 分,共 30 分)

21. 结构化数据与非结构化数据的区别是什么?

22. 一个银行有上亿个储户,如果银行希望找到最高的存储金额是多少,结合下图,说明基于 MapReduce 模型的寻找最大值的过程。



23. 什么是数据质量?简述数据质量的四要素。

24. 什么是特征约简?

25. 什么是聚类?聚类与分类有何不同。

26. 请例举 5 种大数据可视分析技术。

得　分	评卷人

四、应用题(每小题 10 分,共 20 分)

27. 根据图中所示网络爬虫工作原理,说明①～⑤的含义。

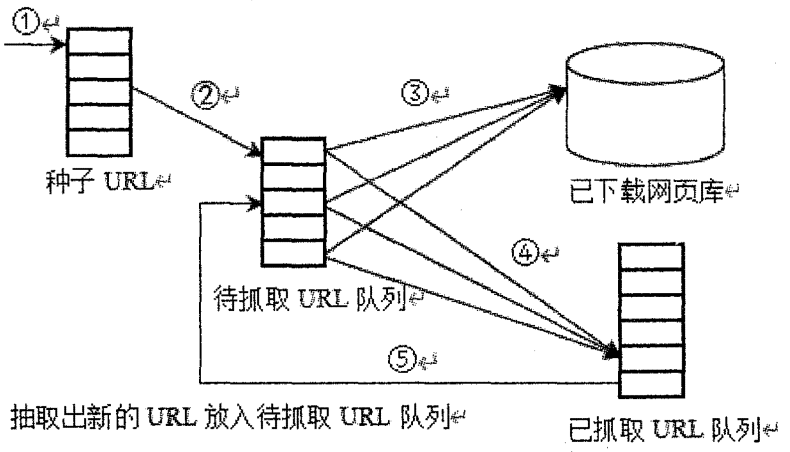


图　网络爬虫工作原理

28. 假设有 9、24、15、41、7、10、18、67、25 共 9 个数，分为 3 箱，各箱的数据分配如下：

箱 1：9、24、15

箱 2：41、7、10

箱 3：18、67、25

(1)按箱平均值法，求箱 1 的平滑数据值。

(2)按箱中值法，求箱 2 的平滑数据值。

(3)按箱边界值法，求箱 3 的平滑数据值。

试卷代号:1480

国家开放大学2020年春季学期期末统一考试

大数据技术导论　试题答案及评分标准

（供参考）

2020年9月

一、选择题(选择一个正确答案的代码填入括号中,每小题3分,共30分)

1. C	2. A	3. C	4. D	5. A
6. D	7. C	8. C	9. C	10. D

二、判断题(正确的划√,错误的划×,每小题2分,共20分)

11. √	12. √	13. ×	14. √	15. ×
16. √	17. √	18. ×	19. ×	20. ×

三、简答题(每小题5分,共30分)

21. 结构化数据与非结构化数据的区别是什么?

答:结构化数据是具有数据结构描述信息的数据,也就是说,结构化数据是先有结构,后有数据,例如各类表格是结构化数据(2分);而非结构化数据是无固定结构来表现的数据,也就是说,只有数据,无结构,例如图形、图像、音频和视频等(3分)。

22. 一个银行有上亿个储户,如果银行希望找到最高的存储金额是多少,结合下图,说明基于MapReduce模型的寻找最大值的过程。

答:首先将数字分布存储在不同块中,以某几个块为一个Map,找出各个Map中最大的值(3分),例如最左列为1000,最右列为985,然后将每个Map中的最大值做Reduce操作,即找出最大值1000后输出(2分)。

23. 什么是数据质量?简述数据质量的四要素。

答:数据质量是数据适合使用的程度,也是数据满足特定用户期望的程度。(1分)

数据质量的四要素为:数据的准确性、数据的完整性、数据的一致性和数据的及时性。

(各1分)

24. 什么是特征约简？

答：特征约简是在保留、提高原有判别能力的前提下，从原有的特征中删除不重要或不相关的特征，或者通过对特征进行重组来减少特征的个数，同时减少特征向量的维度(3 分)。也就是说，特征约简的输入是一组特征，输出也是一组特征，但是输出特征是输入特征的子集(2 分)。

25. 什么是聚类？聚类与分类有何不同。

答：聚类就是自动将数据对象分成多个类或簇，划分的原则是在同一个簇中的数据对象具有较高的相似度，而不同簇中的数据对象相似度差别较大(3 分)。聚类与分类不同的是，聚类操作中要划分的类事先未知，类的形成完全是由数据驱动，属于一种无指导的学习方法(2 分)。

26. 请例举 5 种大数据可视分析技术。

答：大数据可视分析技术有：原位交互分析技术；数据存储技术；可视分析算法；数据移动、传输和网络架构；不确定性的量化；并行计算；面向领域与开发的库、框架以及工具；用户界面与交互设计。(每一点 1 分，有 5 点即满分)

四、应用题(每小题 10 分，共 20 分)

27. 解：(每一条 2 分)

①首先人工选取一部分种子 URL；

②将这些 URL 放入待抓取 URL 队列；

③从待抓取 URL 队列中取出待抓取 URL，解析 DNS 得到主机 IP，并将 URL 对应的网页下载下来，存储到自己的网页库中。

④将这些已抓取的 URL 放入已抓取 URL 队列中；

⑤分析已抓取网页中的其他 URL，并将 URL 放入待抓取的 URL 队列中，进行下一个循环。

28. 解：箱 1：16、16、16(3 分)

箱 2：10、10、10(3 分)

箱 3：18、25、25(4 分)