一、选择题

1

可视分析是一种（）？

【A.】交互式用户界面模型

【B.】结构化用户界面模型

【C.】非结构化用户界面模型

【D.】独立式用户界面模型

【答案】A

2

数据科学的研究对象是（）？

【A.】药品成分

【B.】文学作品

【C.】数据界的数据

【D.】人类历史

【答案】C

3

下面不是研究数据方法的是（）？

【A.】统计学

【B.】机器学习

【C.】心理分析

【D.】数据挖掘

【答案】C

4

下面是数据科学的主要研究内容的是（）？

【A.】数据泄露

【B.】数据交易

【C.】数据盗窃

【D.】数据分析

【答案】D

5

下面不是数据产品开发的特征的是（）？

【A.】多样性

【B.】无竞争性

【C.】增值性

【D.】层次性

【答案】B

6

下面不属于互联网大数据的是（）？

【A.】视频

【B.】图片

【C.】音频

【D.】心情

【答案】D

7

下面不属于大数据的特性是（）？

【A.】数据量大

【B.】多样性

【C.】真实性差

【D.】具有价值

【答案】C

8

下面属于结构化数据的是（）？

【A.】表格数据

【B.】图形

【C.】图像

【D.】HTML文档

【答案】A

9

下面不属于大数据的处理过程的是（）？

【A.】数据获取、储存

【B.】数据清洗

【C.】数据分析

【D.】数据安全

【答案】D

10

下面不属于大数据离线处理特点的是（）？

【A.】数据保存时间短

【B.】数据不会发生改变

【C.】可进行复杂的批量计算

【D.】方便查询计算结果

【答案】A

11

下面不属于大数据处理模式的是（）？

【A.】离线处理

【B.】在线处理

【C.】手动计算

【D.】交互处理

【答案】C

12

下面仅属于大数据在线处理模式的是（）？

【A.】批量计算

【B.】流式计算

【C.】手动计算

【D.】交互处理

【答案】B

13

下面不属于流式数据源的是（）？

【A.】硬盘数据

【B.】传感器数据

【C.】图像数据

【D.】流媒体传输

【答案】A

14

下面不属于流式数据的特点的是（）？

【A.】实时性

【B.】层次性

【C.】易失性

【D.】突发性

【答案】B

15

下面不属于大数据计算模式的类型的是（）？

【A.】批量计算

【B.】手动计算

【C.】流式计算

【D.】交互式计算

【答案】B

16

下面不属于数据流技术应用的方式是（）？

【A.】指定查询

【B.】模糊查询

【C.】即席计算

【D.】中间计算

【答案】B

17

下面不属于判断大数据应用成功的指标是（）？

【A.】创造价值

【B.】具备高速度

【C.】在本质上提高

【D.】低性能

【答案】D

18

基于任务的定义和分类，下列不属于可视分析关注点的是（）？

【A.】以用户价值为关注点

【B.】以用户意图为关注点

【C.】以同户行为为关注点

【D.】以软件操作为关注点

【答案】A

19

下面不属于企业大数据应用成功的考虑因素是（）？

【A.】成本

【B.】客户满意度

【C.】附加收益

【D.】工人满意度

【答案】D

20

下面不属于大数据在通讯行业的应用是（）？

【A.】预测客户行为

【B.】市场监控

【C.】智能电表

【D.】市场预警

【答案】C

21

MapReduce以什么方式进行分布式计算（）？

【A.】文本

【B.】函数

【C.】数据集

【D.】矩阵

【答案】B

22

在Hadoop中，将每一次计算请求称为一个（）？

【A.】文本

【B.】函数

【C.】数据集

【D.】作业

【答案】D

23

在Hadoop中，将每个作业拆分为若干个（）？

【A.】文本

【B.】Map任务

【C.】数据集

【D.】Reduce任务

【答案】B

24

在Hadoop中，作业服务器被称为（）？

【A.】Job

【B.】Map

【C.】Master

【D.】Reduce

【答案】C

25

下列不属于一个作业的计算流程的是（）？

【A.】作业开始

【B.】作业修改

【C.】Map任务分配

【D.】Reduce任务执行

【答案】B

26

下列不属于Map任务的执行的子步骤的是（）？

【A.】输入准备

【B.】输入修改

【C.】算法执行

【D.】输出生成

【答案】B

27

在MapReduce模型编程中，不属于其步骤的是（）？

【A.】遍历数据

【B.】映射键值

【C.】数据分组

【D.】数据插补

【答案】D

28

Hadoop的核心是（ ）和MapReduce。

【A.】NTFS

【B.】GFS

【C.】HDFS

【D.】FS

【答案】C

29

下列不属于单词计数的Map过程的是（）？

【A.】按列分割文件

【B.】分词处理

【C.】按行分割文件

【D.】排序与合并

【答案】A

30

WordCount完成的功能是统计（）？

【A.】按列分割文件数

【B.】输入文件中单词频数

【C.】按行分割文件数

【D.】输入文件中单词频率

【答案】B

31

Hadoop开发环境部署内容不包括（）？

【A.】安装SSH协议

【B.】Hadoop系统部署

【C.】伪分布式Hadoop环境部署

【D.】电脑断网

【答案】D

32

Reduce任务目标是（）？

【A.】输入文件

【B.】输入数据

【C.】输入图表

【D.】将若干个Map任务生成的中间文件汇总到最后的输出文件

【答案】D

33

Hadoop作业服务器不负责（）？

【A.】接受用户提交的作业

【B.】任务的分配

【C.】执行具体的任务

【D.】管理所有的任务服务器

【答案】C

34

下列不属于文本数据语义特征的是（）？

【A.】词频

【B.】逻辑结构

【C.】动态演化规律

【D.】数据结构

【答案】D

35

文本数据是典型的（）？

【A.】结构化数据

【B.】半结构化数据

【C.】非结构化数据

【D.】混合结构数据

【答案】C

36

MapReduce模型将数据集的大规模操作分发给网络上的各节点，每个节点将已完成的工作和状态更新，周期性地报告给（）。

【A.】作业

【B.】任务

【C.】作业服务器

【D.】任务服务器

【答案】C

37

MapReduce模型的最大优点体现在（）？

【A.】输入方便

【B.】分布式处理

【C.】各类型数据输入

【D.】算法简单

【答案】B

38

MapReduce模型适用场景的特点（）？

【A.】数据可实时处理

【B.】任务不可分解

【C.】任务可分解

【D.】任务需要顺序执行

【答案】C

39

MapReduce将计算过程分解的最大好处是（）？

【A.】复杂化

【B.】快捷化

【C.】并行化

【D.】低成本

【答案】C

40

下列不属于MapReduce模型的限制因素的是（）？

【A.】任务要求

【B.】数据不可无限分割

【C.】通信开销

【D.】集群规模

【答案】A

41

下列不属于信息可视化的分类的是（）？

【A.】一维信息可视化

【B.】三维信息可视化

【C.】截面数据可视化

【D.】时序信息可视化

【答案】C

42

下列属于传统数据获取的特点的是（）？

【A.】数据源多种多样

【B.】数据量巨大

【C.】结构单一

【D.】结构化、半结构化和非结构化数据

【答案】C

43

下列不属于大数据获取的特点的是（）？

【A.】数据源多种多样

【B.】数据量巨大

【C.】结构单一

【D.】结构化、半结构化和非结构化数据

【答案】C

44

下列属于大数据获取的特点的是（）？

【A.】数据源单一

【B.】数据量较小

【C.】结构单一

【D.】包括结构化、半结构化和非结构化数据

【答案】D

45

气泡图由什么演化而来是（）？

【A.】直方图

【B.】条形图

【C.】散点图

【D.】折线图

【答案】C

46

下列不属于文本数据的是（）？

【A.】广告

【B.】音频

【C.】杂志

【D.】图书

【答案】B

47

下列不属于文本数据的获取特点的是（）？

【A.】灵活度高

【B.】速度快

【C.】按需获取

【D.】易处理

【答案】D

48

下列属于意见建议类用户反馈的特点的是（）？

【A.】反馈量大

【B.】反馈速度快

【C.】针对性强

【D.】传播快

【答案】C

49

下列不属于空间数据的特性的是（）？

【A.】定位

【B.】定量

【C.】定性

【D.】时空关系

【答案】B

50

下列不属于网站内部数据库中的数据的是（）？

【A.】用户信息数据

【B.】野外实测数据

【C.】网站产品数据

【D.】网站运营数据

【答案】B

51

下列不属于网站数据采集的数据的是（）？

【A.】非结构化数据

【B.】半结构化数据

【C.】结构化数据

【D.】实验测试数据

【答案】D

52

下列不属于网络爬虫工作过程的是（）？

【A.】获取网页

【B.】修改网页

【C.】解析网页

【D.】储存数据

【答案】B

53

下列不属于通用网络爬虫的特点及要求的是（）？

【A.】数量巨大

【B.】范围较小

【C.】存储空间要大

【D.】爬行速度要快

【答案】B

54

下列不属于通用网络爬虫的结构的是（）？

【A.】页面爬行模块

【B.】页面分析模块

【C.】页面数据库

【D.】页面修改模块

【答案】D

55

下列不属于网页的爬虫策略的是（）？

【A.】深度优先搜索策略

【B.】精确选择搜索策略

【C.】广度优先搜索策略

【D.】最佳优先搜索策略

【答案】B

56

下列不属于通用网络爬虫的局限性的是（）？

【A.】无用网页较多

【B.】对非结构化数据获取相对容易

【C.】难以支持基于语义信息的查询

【D.】有限的网络爬虫服务器资源与无限的网络数据资源之间的冲突

【答案】B

57

下列属于聚焦网络爬虫的特点的是（）？

【A.】数量巨大

【B.】范围较大

【C.】精准筛选

【D.】爬行速度慢

【答案】C

58

下列属于浅聚焦网络爬虫的核心特点的是（）？

【A.】数量巨大

【B.】范围较大

【C.】精准筛选

【D.】选定URL种子

【答案】D

59

下列不属于爬虫对抓取目标的定义原则的是（）？

【A.】目标网页特征

【B.】目标数据量

【C.】目标数据模式

【D.】领域概念

【答案】B

60

下列不属于网络爬虫的数据抓取方式的是（）？

【A.】预先给定的初始抓取种子样本

【B.】预先给定的网页分类目录和与分类目录对应的种子样本

【C.】通过用户行为确定的抓取目标样例

【D.】自行编写种子样本

【答案】D

61

下列不属于NoSQL数据库的特点的是（）？

【A.】需预定义数据模型

【B.】支持透明横向扩展

【C.】将数据进行分区

【D.】保证最终一致性

【答案】A

62

下列不属于NoSQL数据库的存储方式的是（）？

【A.】行式存储

【B.】键值式存储

【C.】图形式存储

【D.】文档式存储

【答案】A

63

下列不属于NoSQL数据库的键值式存储的常见形式的是（）？

【A.】临时型

【B.】附加型

【C.】混合型

【D.】永久型

【答案】B

64

下列不属于数据库架构混合应用模式的是（）？

【A.】OldSQL＋NewSQL混合模式

【B.】OldSQL＋NoSQL+ NewSQL混合模式

【C.】OldSQL＋NoSQL混合模式

【D.】NewSQL＋NoSQL混合模式

【答案】B

65

下列不属于大数据抽取方式的是（）？

【A.】同构同质数据抽取

【B.】异构同质数据抽取

【C.】同构异质数据抽取

【D.】文件型数据抽取

【答案】B

66

下列不属于增量数据抽取特点的是（）？

【A.】抽取发生变化的数据

【B.】快捷

【C.】处理量更多

【D.】需要与数据装载时的更新策略相对应

【答案】C

67

下列不属于地域空间可视化展现的功能的是（）？

【A.】认识功能

【B.】模拟功能

【C.】载负功能

【D.】学习功能

【答案】D

68

时间戳是能表示一份数据在某个特定时间之前已经存在的、完整的、可验证的一个数据，其通常是（）？

【A.】一个数据集

【B.】一个字符序列

【C.】一个日志文件

【D.】一段储存代码

【答案】B

69

下列不属于基于时间戳的增量数据抽取方式的优点的是（）？

【A.】性能优异

【B.】系统设计清晰

【C.】数据抽取简单

【D.】对业务系统无其他要求

【答案】D

70

下列不属于MD5（消息摘要算法）的特点的（）？

【A.】对源系统的倾入性较小

【B.】被动地进行全表数据的比对

【C.】性能优异

【D.】准确性不稳定

【答案】C

71

下列不属于数据清洗的是（）？

【A.】检查数据的完整性

【B.】修改数据

【C.】填补数据

【D.】消除重复

【答案】B

72

下列不属于数据质量的要素的是（）？

【A.】准确性

【B.】完整性

【C.】异质性

【D.】及时性

【答案】C

73

下列不属于数据质量一致性的是（）？

【A.】数据编码一致性

【B.】数据大小一致性

【C.】指标统计一致性

【D.】指标计算一致性

【答案】B

74

【答案】D

75

下列不属于数据清洗算法标准的是（）？

【A.】返回率

【B.】错误返回率

【C.】正确返回率

【D.】精确度

【答案】C

76

下列不会造成文本记录重复相似的是（）？

【A.】插入

【B.】交换

【C.】等价表述

【D.】删除冗余

【答案】D

77

下列不属于文本相似度计算的应用领域的是（）？

【A.】信息检索

【B.】数据挖掘

【C.】文档修改

【D.】机器翻译

【答案】C

78

下列不属于文本相似度计算流程的是（）？

【A.】文本排序

【B.】文本分词

【C.】统计词频

【D.】写出词频向量

【答案】A

79

增量数据抽取方式只抽取（）？

【A.】PB级数据

【B.】不变数据

【C.】变化数据

【D.】有价值数据

【答案】C

80

数据清洗算法的衡量标准主要包含 （ ）、错误返回率和精确度。

【A.】冗余度

【B.】返回率

【C.】可用性

【D.】一致性

【答案】B

81

下列不属于数据转换中适合对数转换的数据是（）？

【A.】部分正偏态数据

【B.】等比数据

【C.】等差数据

【D.】数值相差不大的数据

【答案】C

82

下列不属于数据转换中适合平方根转换的数据是（）？

【A.】泊松分布数据

【B.】轻度偏态数据

【C.】样本方差和平均数呈现正相关的数据

【D.】数值相差不大的数据

【答案】D

83

噪声是指测量数据中的（）？

【A.】随机误差

【B.】均值

【C.】方差

【D.】极差

【答案】A

84

下列不属于地图的构成要素的是（）？

【A.】图形要素

【B.】价格要素

【C.】数学要素

【D.】辅助要素

【答案】B

85

下列不属于数据平滑方法的是（）？

【A.】移动平均法

【B.】最大最小值标准化

【C.】指数平滑法

【D.】分箱平滑法

【答案】B

86

下列不属于移动平均法的是（）？

【A.】一次移动平均法

【B.】二次移动平均法

【C.】混合移动平均法

【D.】多次移动平均法

【答案】C

87

下列属于一次移动平均法的适用范围的是（）？

【A.】水平变动的时间序列数据

【B.】长期变动趋势的时间序列数据

【C.】循环性变动的时间序列数据

【D.】季节变动趋势的时间序列数据

【答案】A

88

下列属于二次移动平均法的适用范围的是（）？

【A.】水平变动的时间序列数据

【B.】长期变动趋势的时间序列数据

【C.】循环性变动的时间序列数据

【D.】直线上升或下降趋势的时间序列数据

【答案】D

89

下列不属于指数平滑法的适用范围的是（）？

【A.】稳定变化态势的时间序列数据

【B.】不稳定变化的时间序列数据

【C.】规则变化的时间序列数据

【D.】直线上升或下降趋势的时间序列数据

【答案】B

90

下列不属于指数平滑法的适用范围的是（）？

【A.】近期趋势预测

【B.】短期趋势预测

【C.】中期趋势预测

【D.】长期趋势预测

【答案】D

91

下列不属于分箱平滑法参照的是（）？

【A.】箱内数值的均值

【B.】箱内数值的中值

【C.】箱内数值的边界值

【D.】箱内数值的方差

【答案】D

92

数据规范化是将原来的度量值转换为（）？

【A.】无量纲的值

【B.】原来度量值的均值

【C.】原来度量值的方差

【D.】原来度量值的极差

【答案】A

93

下列不属于数据规范化的作用的是（）？

【A.】加快学习速度

【B.】规范事物概念

【C.】避免单位的影响

【D.】呈现数据的集中程度

【答案】D

94

假定某属性ｘ的最小值、最大值分别为12000和98000，将属性ｘ映射到 ［0, 1］中，根据最小最大规范化方法，ｘ的值 73600（设定值）将转换为（）？

【A.】0.716

【B.】0.616

【C.】0.516

【D.】0.816

【答案】A

95

Z分数规范化方法依据的是原始数据的（）？

【A.】均值和中位数

【B.】中位数和极差

【C.】极差和方差

【D.】均值和标准差

【答案】D

96

如果ｘ的平均值和标准差分别为 54000和 16000。使用ｚ分数规范化方法，ｘ的值 73600被转换为（）？

【A.】1.125

【B.】1.025

【C.】1.225

【D.】1.325

【答案】C

97

Z分数规范化方法中对于离群点，均值绝对偏差相比标准差（）？

【A.】更有效

【B.】更平稳

【C.】更精确

【D.】更鲁棒

【答案】D

98

小数定标规范化方法中，其小数点移动的位数依赖于（）？

【A.】最大绝对值

【B.】最小绝对值

【C.】均值绝对值

【D.】中位数绝对值

【答案】A

99

数据平滑方法主要有指数平滑法、移动平均法和（）？

【A.】统计法

【B.】最短距离法

【C.】分箱平滑法

【D.】聚类方法

【答案】C

100

数据规范方法主要有最小最大规范法、z分数规范法和（）？

【A.】极差规范法

【B.】小数定标规范法

【C.】模糊规范法

【D.】聚类方法

【答案】B

101

下列不属于数据约简策略的是（）？

【A.】特征约简

【B.】样本约简

【C.】数值约简

【D.】文本约简

【答案】D

102

下列不属于特征约简的步骤的是（）？

【A.】搜索过程

【B.】填补过程

【C.】评估过程

【D.】分类过程

【答案】B

103

【答案】B

104

系统抽样又称之为（）？

【A.】随机抽样

【B.】等距抽样

【C.】分层抽样

【D.】类型抽样

【答案】B

105

下列不属于影响突出对比的可视化展现是（）？

【A.】比例选择

【B.】颜色使用

【C.】图形形状

【D.】图形内容

【答案】D

106

下列不属于检验假设方法的是（）？

【A.】提出假设

【B.】选择统计量

【C.】修正统计量

【D.】根据显著性水平进行判断

【答案】C

107

分层抽样又称之为（）？

【A.】随机抽样

【B.】等距抽样

【C.】系统抽样

【D.】类型抽样

【答案】D

108

数据立方体是一类多维矩阵，让用户从多个角度探索和分析数据集，通常是一次同时考虑几个维度（）？

【A.】一个

【B.】两个

【C.】三个

【D.】四个

【答案】C

109

三维的数据立方体看作是一组类似的互相叠加起来的（）？

【A.】一维表格

【B.】二维表格

【C.】三维表格

【D.】四维表格

【答案】B

110

【答案】C

111

维数灾难通常是指在涉及（）计算的问题中，随着维数的增加，计算量呈指数倍增长的 一种现象。

【A.】数字

【B.】文本

【C.】向量

【D.】矩阵

【答案】C

112

下列不属于维数灾难问题的特点（）？

【A.】计算量巨大

【B.】成本降低

【C.】结果不理想

【D.】无法反映数据的本质特征

【答案】B

113

下列不属于数据约简的特点（）？

【A.】计算量巨大

【B.】成本降低

【C.】较少存储量

【D.】方便分类

【答案】A

114

下列不属于数据约简的分类基准的是（）？

【A.】约简维数的大小

【B.】成本的大小

【C.】数据时序

【D.】有无监督信息

【答案】B

115

下列不属于数据约简中有无监督信息分类的是（）？

【A.】监督式维数约简

【B.】半监督式维数约简

【C.】非监督式维数约简

【D.】混合式维数约简

【答案】D

116

下列不属于无参数值约简技术的是（）？

【A.】直方图

【B.】聚类

【C.】选择

【D.】线性回归模型

【答案】D

117

下列不属于信息孤岛带来的问题的是（）？

【A.】数据不能共享

【B.】冗余数据

【C.】数据优化

【D.】垃圾数据

【答案】C

118

数据集成最复杂和困难的任务是（）？

【A.】数据插补

【B.】数据格式转换

【C.】数据修改

【D.】冗余数据处理

【答案】B

119

下列不属于数据集成的核心问题是（）？

【A.】有效性

【B.】异构性

【C.】分布性

【D.】自治性

【答案】A

120

下列不属于数据集成的构成方式的是（）？

【A.】联邦数据库集成模式

【B.】中间件集成模式

【C.】数据仓库集成模式

【D.】地区数据库集成模式

【答案】D

121

大数据需要分析的是（）？

【A.】发展趋势

【B.】发展现状

【C.】发展模式

【D.】发展制度

【答案】A

122

大数据分析的核心是（）？

【A.】数据收集

【B.】数据管理

【C.】数据挖掘

【D.】数据交易

【答案】C

123

下列不属于数据挖掘的特点的是（）？

【A.】数据量大

【B.】目标清晰

【C.】处理不同类型的数据

【D.】结果不易解释

【答案】B

124

下列属于数据分析的特点的是（）？

【A.】数据量大

【B.】目标不清晰

【C.】处理不同类型的数据

【D.】结果不易解释

【答案】A

125

从分析的结果上看，大数据分析主要分为（）？

【A.】探索性数据分析、证实性数据分析、定性数据分析

【B.】证实性数据分析、定性数据分析、离线数据分析

【C.】探索性数据分析、定性数据分析、离线数据分析

【D.】探索性数据分析、证实性数据分析、离线数据分析

【答案】A

126

从分析方式上看，大数据分析主要分为（）？

【A.】在线数据分析、离线数据分析、交互式数据分析

【B.】交互式数据分析、在线数据分析、定性数据分析

【C.】定性数据分析、离线数据分析、交互式数据分析

【D.】离线数据分析、在线数据分析、定性数据分析

【答案】A

127

探索性数据分析是基于什么角度来说明数据分析方法的（）？

【A.】数据本身

【B.】数据范围

【C.】模型假设

【D.】统计推断

【答案】A

128

下列不属于探索性数据分析的特点的是（）？

【A.】分离出数据的模式

【B.】分离出数据的特点

【C.】揭示数据对模型的偏离

【D.】计算模型对数据的拟合度

【答案】D

129

下列不属于探索性数据分析的内容的是（）？

【A.】检查数据错误

【B.】填补数据缺失

【C.】获取数据分布特征

【D.】观察数据规律

【答案】B

130

下列不属于数据位置特征的是（）？

【A.】均值

【B.】中位数

【C.】方差

【D.】四分位数

【答案】C

131

下列不属于数据分散性特征的是（）？

【A.】均值

【B.】极差

【C.】方差

【D.】变异系数

【答案】A

132

统计学是收集、分析、表述和（）的科学？

【A.】整理数据

【B.】计算数据

【C.】填补数据

【D.】解释数据

【答案】D

133

统计的基础是（）？

【A.】统计工作

【B.】计算数据

【C.】统计数据

【D.】统计科学

【答案】A

134

下列不属于模型的组成的是（）？

【A.】目标

【B.】数据

【C.】变量

【D.】关系

【答案】B

135

下列不属于相关系数的是（）？

【A.】简单相关系数

【B.】复相关系数

【C.】典型相关系数

【D.】常态相关系数

【答案】D

136

下列不属于相关分析的是（）？

【A.】线性相关分析

【B.】复相关分析

【C.】偏相关分析

【D.】距离分析

【答案】B

137

下列不属于回归分析的步骤的是（）？

【A.】确定自变量与因变量

【B.】对变量进行分类

【C.】根据变量进行回归建模

【D.】模型检验

【答案】B

138

下列不属于回归分析的类别的是（）？

【A.】一元回归分析

【B.】二元回归分析

【C.】多元回归分析

【D.】线性回归分析

【答案】B

139

下列不属于判别分析的类别的是（）？

【A.】两组判别法

【B.】多组判别法

【C.】逐步判别法

【D.】面积判别法

【答案】D

140

下列不属于判别方法的是（）？

【A.】最大似然法

【B.】最小二乘法

【C.】距离判别法

【D.】贝叶斯判别法

【答案】B

141

Z分数规范化方法中，下列可以替换标准差的是（）？

【A.】均值

【B.】方差

【C.】极差

【D.】平均值绝对偏差

【答案】D

142

下列不属于确定样本子集大小的因素的是（）？

【A.】计算成本

【B.】存储要求

【C.】估计量的范围

【D.】估计量的精度

【答案】C

二、判断题

1

计算机科学是算法与算法变换的科学。（）

【A.】√

【B.】 ×

【答案】A

计算机科学是算法与算法变换的科学。

2

数据科学是通过科学方法探索数据，以获得有价值的发现。（）

【A.】√

【B.】 ×

【答案】A

数据科学是通过科学方法探索数据，以获得有价值的发现。

3

数据科学的发展不仅可以推动学科的发展，而且能够助推相关产业的发展与进步。（）

【A.】√

【B.】 ×

【答案】A

数据科学的发展不仅可以推动数学、计算机科学、人工智能、统计学、天体信息学、生物信息学、计算社会学等学科的发展，而且能够大力助推相关产业的发展与进步。

4

数据科学家不需要具备计算机科学、统计学的知识和应用领域的行业经验。（）

【A.】√

【B.】 ×

【答案】B

数据科学家需要具备计算机科学、统计学的知识和应用领域的行业经验。

5

Cyber空间是指以计算机技术、现代通信网络技术、虚拟现实技术等信息技术的综合运用为基础，以知识和信息为内容的新型空间。（）

【A.】√

【B.】 ×

【答案】A

Cyber空间是指以计算机技术、现代通信网络技术、虚拟现实技术等信息 技术的综合运用为基础，以知识和信息为内容的新型空间。

6

数据能够记录人类的行为，包括工作、生活和社会的发展。（）

【A.】√

【B.】 ×

【答案】A

数据能够记录人类的行为，包括工作、生活和社会的发展。

7

研究数据仅仅包括数据采集和数据分析。（）

【A.】√

【B.】 ×

【答案】B

研究数据主要包括数据采集、数据存储和数据分析。

8

数据加工与传统数据处理的不同之处在于，其更加强调数据处理中的增值过程。（）

【A.】√

【B.】 ×

【答案】A

数据加工与传统数据处理的不同之处在于，其更加强调数据处理中的增值过程。

9

大数据只来自互联网世界。（）

【A.】√

【B.】 ×

【答案】B

大数据主要来自互联网世界与物理世界。

10

数据类型包括结构化数据、非结构化数据和半结构化数据。（）

【A.】√

【B.】 ×

【答案】A

数据类型包括结构化数据、非结构化数据和半结构化数据。

11

网页数据是一种半结构化数据。（）

【A.】√

【B.】 ×

【答案】A

网页数据就是一种典型的半结构化数据。

12

批量计算是一种在线计算，支持海量作业 并发规模，系统自动完成资源管理、作业调度和数据加载，并按实际使用量计费。（）

【A.】√

【B.】 ×

【答案】B

批量计算是一种离线计算。

13

静态数据是流式数据。（）

【A.】√

【B.】 ×

【答案】B

静态数据不是流式数据。

14

流式计算适用于无须先存储，可以直接进行数据计算，实时性要求很严格，但对数据的精确度要求较宽松的应用场景。（）

【A.】√

【B.】 ×

【答案】A

流式计算适用于无须先存储，可以直接进行数据计算，实时性要求很严格，但对数据的精确度要求较宽松的应用场景。

15

实时计算系统的设计需要考虑高延迟、高性能、分布式、可扩展、高容错。（）

【A.】√

【B.】 ×

【答案】B

实时计算系统的设计需要考虑低延迟、高性能、分布式、可扩展、高容错。

16

流式查询主要有两种方式，一种是指定查询，另一种是即席查询。（）

【A.】√

【B.】 ×

【答案】A

流式查询主要有两种方式，一种是指定查询，另一种是即席查询。

17

流式数据的处理是批量处理。（）

【A.】√

【B.】 ×

【答案】B

流式数据的处理也是连续处理，而不是批量处理。

18

流式计算不宜用持久稳定关系建模，而适用瞬态数据流建模。（）

【A.】√

【B.】 ×

【答案】A

流式计算不宜用持久稳定关系建模，而适用瞬态数据流建模。

19

为了更有效地处理数据，应该尽可能地限制静态数据。（）

【A.】√

【B.】 ×

【答案】A

为了有效地处理数据，人们应该尽可能地限制静态数据。

20

大数据的无序性是指各数据流之间无序，而同一数据流内部各数据元素之间是有序的。（）

【A.】√

【B.】 ×

【答案】B

大数据的无序性是指各数据流之间无序，而同一数据流内部各数据元素之间也无序。

21

MapReduce由 Map和 Reduce两个阶段组成,用户只需要编写 Map和 Reduce两个函数就可以完成简单的分布式程序的设计。（）

【A.】√

【B.】 ×

【答案】A

MapReduce由 Map和 Reduce两个阶段组成,用户只需要编写 Map和 Reduce两个函数就 可以完成简单的分布式程序的设计。

22

Map相对独立且并行运行，对存储系统中的文件按列处理，并产生键值对。（）

【A.】√

【B.】 ×

【答案】B

Map相对独立且并行运行，对存储系统中的文件按行处理，并产生键值对。

23

在Hadoop中，将每一次计算请求称为一个作业。（）

【A.】√

【B.】 ×

【答案】A

在Hadoop中，将每一次计算请求称为一个作业。

24

与分布式文件系统相比，MapReduce框架可定制性强。（）

【A.】√

【B.】 ×

【答案】A

与分布式文件系统相比，MapReduce框架还有一个特点就是可定制性强。

25

通常一个作业的输入都是基于分布式文件系统的文件，而对于一个 Map任务而言，它的输入是输入文件的一个数据块，或者是数据块的一部分，也可跨越数据块。（）

【A.】√

【B.】 ×

【答案】B

通常一个作业的输入都是基于分布式文件系统的文件，而对于一个 Map任务而言，它的输入是输入文件的一个数据块，或者是数据块的一部分，但通常不跨越数据块。

26

Reduce任务与 Map任务的最大不同是 Map任务的文件都存储于本地，而 Reduce任务需要到多处采集。（）

【A.】√

【B.】 ×

【答案】A

Reduce任务与 Map任务的最大不同是 Map任务的文件都存储于本地，而 Reduce任务需要到多处采集。

27

MapReduce模型适用场景的特点是任务可被分解成相互联系的子问题。（）

【A.】√

【B.】 ×

【答案】B

MapReduce模型适用场景的特点是任务可被分解成相互独立的子问题。

28

可视分析是一种独立式的图形用户界面模型。（）

【A.】√

【B.】 ×

【答案】B

可视分析是一种交互式的图形用户界面模型。

29

在单词计数的Map过程中需要将文件进行按列分割。（）

【A.】√

【B.】 ×

【答案】B

在单词计数的Map过程中需要将文件进行按行分割。

30

在单词计数中，Reduce方法的输入参数 key为单个单词，而 value是由各 Mapper类上对应单词的计数值所组成的列表，所以只要遍历 value并求和，即可得到某个单词出现的总次数。（）

【A.】√

【B.】 ×

【答案】A

在单词计数中，Reduce方法的输入参数 key为单个单词，而 value是由各 Mapper类上对应单词的计数值所组成的列表，所以只要遍历 value并求和，即可得到某个单词出现的总次数。

31

WordCount完成的功能是统计输入文件中的每个单词出现的频率（）

【A.】√

【B.】 ×

【答案】B

WordCount完成的功能是统计输入文件中的每个单词出现的次数。

32

在MapReduce程序执行过程中，用户程序中的MapReduce类库首先将输入文档进行分割，用户也可以通过设置参数对其大小进行控制（）

【A.】√

【B.】 ×

【答案】A

在MapReduce程序执行过程中，用户程序中的MapReduce类库首先将输入文档进行分割，用户也可以通过设置参数对其大小进行控制。

33

在MapReduce程序执行过程中，由Master负责分配任务，分配的原则是 Master选择空闲的 Worker并为其分配一个 Map任务或一个 Reduce任务（）

【A.】√

【B.】 ×

【答案】A

在MapReduce程序执行过程中，由Master负责分配任务，分配的原则是 Master选择空闲的 Worker并为其分配一个 Map任务或一个 Reduce任务。

34

网站外部数据主要包括互联网环境数据、竞争对手数据、合作伙伴数据和用户数据等。（）

【A.】√

【B.】 ×

【答案】A

网站外部数据主要包括互联网环境数据、竞争对手数据、合作伙伴数据和用户数据等。

35

在MapReduce程序执行过程中，Map和Reduce函数接收的都是键值对。（）

【A.】√

【B.】 ×

【答案】A

在MapReduce程序执行过程中，Map和Reduce函数接收的都是键值对。

36

MapReduce模型通过将数据集的大规模操作分发给网络上的各节点，每个节点将已完成的工作和状态更新，周期性地报告给 Worker。（）

【A.】√

【B.】 ×

【答案】B

MapReduce模型通过将数据集的大规模操作分发给网络上的各节点，每个节点将已完成 的工作和状态更新，周期性地报告给 Master。

37

MapReduce的基本原理就是将大数据分成小块逐个分析，最后将提取出来的数据汇总分析，进而获得需要的结果。（）

【A.】√

【B.】 ×

【答案】A

MapReduce的基本原理就是将大数据分成小块逐个分析，最后将提取出来的数据汇总分析，进而获得需要的结果。

38

MapReduce模型中，Map针对每一个输入元素都要生成一个输出元素，Reduce针对每一个输入列表都要生成一个输出元素。（）

【A.】√

【B.】 ×

【答案】A

MapReduce模型中，Map针对每一个输入元素都要生成一个输出元素，Reduce针对每一个输入列表都要生成一个输出元素。

39

通常情况下如果集群的规模在百个节点以上，MapReduce的速度可以和节点的数目成正比。（）

【A.】√

【B.】 ×

【答案】A

通常情况下如果集群的规模在百个节点以上，MapReduce的速度可以和节点的数目成正比。

40

由于集群模式和单节点模式运行 Hadoop系统都需要使用 SSH登录，因此在安装 Hadoop 系统之前，首先需要安装配置 SSH协议。（）

【A.】√

【B.】 ×

【答案】A

由于集群模式和单节点模式运行 Hadoop系统都需要使用 SSH登录，因此在安装 Hadoop 系统之前，首先需要安装配置 SSH协议。

41

获取的数据是指已被转换为电信号的各种物理量，如温度、水位、风速、压力等。（）

【A.】√

【B.】 ×

【答案】A

获取的数据是指已被转换为电信号的各种物理量，如温度、水位、风速、压力等。

42

大数据的获取要避免重复数据。（）

【A.】√

【B.】 ×

【答案】A

大数据的获取要避免重复数据。

43

在互联网营销中，用户反馈承担的核心任务是为产品收集用户舆情信息。（）

【A.】√

【B.】 ×

【答案】A

在互联网营销中，用户反馈承担的核心任务是为产品收集用户舆情信息。

44

图像数字化是进行数字图像处理的前提。（）

【A.】√

【B.】 ×

【答案】A

图像数字化是进行数字图像处理的前提。

45

图像数字化是将连续色调的模拟图像经采样量化后转换成数字影像的过程。（）

【A.】√

【B.】 ×

【答案】A

图像数字化是将连续色调的模拟图像经采样量化后转换成数字影像的过程。

46

图像信息获取的方法只有扫描技术。（）

【A.】√

【B.】 ×

【答案】B

图像信息获取的主要方法是扫描技术，另一种方法是直接运用数字摄影技术。

47

图形数字化是将图形的连续模拟量转换成离散的数字量的过程。（）

【A.】√

【B.】 ×

【答案】A

图形数字化是将图形的连续模拟量转换成离散的数字量的过程。

48

空间数据是一种用点、线、面以及实体等基本空间数据结构来表示自然世界的数据。（）

【A.】√

【B.】 ×

【答案】A

空间数据是一种用点、线、面以及实体等基本空间数据结构来表示自然世界的数据。

49

空间数据的获取不包括对多媒体数据进行获取。（）

【A.】√

【B.】 ×

【答案】B

空间数据获取的任务包括对地图数据、野外实测数据、空间定位数据、摄影测量与遥感图像、多媒体数据等进行获取。

50

网站内部数据是网站最容易获取的数据，其通常存放在网站的文件系统或数据库中，也是与网站自身最为密切相关的数据，是网站分析最常用的数据来源。（）

【A.】√

【B.】 ×

【答案】A

网站内部数据是网站最容易获取的数据，其通常存放在网站的文件系统或数据库中，也是与网站自身最为密切相关的数据，是网站分析最常用的数据来源。

51

网站数据采集只是将网站上的结构化数据从网页中提取出来，并将其存储到统一的本地数据文件中。（）

【A.】√

【B.】 ×

【答案】B

网站数据采集是将网站上的非结构化数据、半结构化数据和结构化数据从网页中提取出来，并将其存储到统一的本地数据文件中。

52

网络爬虫的过程主要分为获取网页、解析网页和存储数据三部分，其是按照一定的获取网页规则，自动地抓取互联网数据的软件。（）

【A.】√

【B.】 ×

【答案】A

网络爬虫的过程主要分为 获取网页、解析网页和存储数据三部分，其是按照一定的获取网页规则，自动地抓取互联网数据的软件。

53

网络爬虫可以分为通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫等。（）

【A.】√

【B.】 ×

【答案】A

网络爬虫可以分为通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫等。

54

网络爬虫从一个或若干初始网页的 URL开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL放入队列，直到满足系统的停止条件为止。（）

【A.】√

【B.】 ×

【答案】A

网络爬虫从一个或若干初始网页的 URL开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL放入队列，直到满足系统的停止条件为止。

55

通用网络爬虫又称为全网爬虫，其可将爬行对象从一些种子 URL扩充到整个 Web，主要为门户站点搜索引擎和大型 Web服务采集数据。（）

【A.】√

【B.】 ×

【答案】A

通用网络爬虫又称为全网爬虫，其可将爬行对象从一些种子 URL扩充到整个 Web，主 要为门户站点搜索引擎和大型 Web服务采集数据。

56

网页的爬行策略可以分为深度优先搜索策略、广度优先搜索策略、最佳优先搜索策略和反向链接数搜索策略，其中深度优先搜索策略是最常使用的方法。（）

【A.】√

【B.】 ×

【答案】B

网页的爬行策略可以分为深度优先搜索策略、广度优先搜索策略、最佳优先搜索策略和反向链接数搜索策略，其中广度优先搜索策略和最佳优先搜索策略是经常使用的方法。

57

聚焦网络爬虫又称为主题爬虫，是面向特定主题的一种网络爬虫程序。（）

【A.】√

【B.】 ×

【答案】A

聚焦网络爬虫又称为主题爬虫，是面向特定主题的一种网络爬虫程序。

58

聚焦网络爬虫与通用网络爬虫的区别之处在于聚焦网络爬虫在实施网页抓取时要进行主题筛选，尽量保证只抓取与主题相关的网页信息。（）

【A.】√

【B.】 ×

【答案】A

聚焦网络爬虫与通用网络爬虫的区别之处在于聚焦网络爬虫在实施网页抓取时要进行主题筛选，尽量保证只抓取与主题相关的网页信息。

59

网页分析算法可以归纳为基于网络拓扑、基于网页内容和基于用户访问行为三种类型。（）

【A.】√

【B.】 ×

【答案】A

网页分析算法可以归纳为基于网络拓扑、基于网页内容和基于用户访问行为三种类型。

60

网络拓补分析算法是基于网页之间的链接，通过已知的网页或数据对与其有直接或间接链接关系的对象作出评价的算法，其又分为网页粒度、网站粒度两种算法。（）

【A.】√

【B.】 ×

【答案】B

网络拓补分析算法是基于网页之间的链接，通过已知的网页或数据对与其有直接或间接链接关系的对象作出评价的算法，其又分为网页粒度、网站粒度以及网页块粒度三种算法。

61

从应用的构建架构角度出发，我们可以将数据库归纳为 OldSQL数据库、NoSQL数据库和 NewSQL数据库。（）

【A.】√

【B.】 ×

【答案】A

从应用的构建架构角度出发，我们可以将数据库归纳为 OldSQL数据库、NoSQL数据库和NewSQL数据库。

62

OldSQL数据库是指传统的关系数据库，NoSQL数据库是指非结构化数据库，而 NewSQL数据库是介于 OldSQL数据库和 NoSQL数据库两者之间的数据库。（）

【A.】√

【B.】 ×

【答案】A

OldSQL数据库是指传统的关系数据库，NoSQL数据库是指非结构化数据库，而 NewSQL数据库是介于 OldSQL数据库和 NoSQL数据库两者之间的数据库。

63

OldSQL数据库适用于数据分析应用，NewSQL数据库适用于事务处理应用，NoSQL数据库适用于互联网应用。（）

【A.】√

【B.】 ×

【答案】B

OldSQL数据库适用于事务处理应用，NewSQL数据库适用于数据分析应用，NoSQL数据库适用于互联网应用。

64

大数据可视分析通过交互可视界面来进行分析、推理和决策，可视分析与各个领域的数据形态、大小及其应用密切相关。（）

【A.】√

【B.】 ×

【答案】A

大数据可视分析通过交互可视界面来进行分析、推理和决策，可视分析与各个领域的数据形态、大小及其应用密切相关。

65

NoSQL主要指非关系型、分布式、不提供 ACID 特性的数据库设计模式。（）

【A.】√

【B.】 ×

【答案】A

NoSQL主要指非关系型、分布式、不提供 ACID 特性的数据库设计模式。

66

NoSQL数据库代表了一系列的、不同类型的相互关联的数据存储与处理的技术的集合。（）

【A.】√

【B.】 ×

【答案】A

NoSQL数据库代表了一系列的、不同类型的相互关联的数据存储与处理的技术的集合。

67

NoSQL数据库与 SQL数据库显著的区别是 NoSQL数据库不使用 SQL作为查询语言，其数据存储不使用固定的表格模式，具有横向可扩展性的特征。（）

【A.】√

【B.】 ×

【答案】A

NoSQL数据库与 SQL数据库显著的区别是 NoSQL数据库不使用 SQL作为查询语言，其数据存储不使用固定的表格模式，具有横向可扩展性的特征。

68

NewSQL数据库是指各种新型的可扩展／高性能数据库，这类数据库不仅具有 NoSQL数据库对海量数据的存储管理能力，还保持了传统数据库的 ACID和 SQL等特性。（）

【A.】√

【B.】 ×

【答案】A

NewSQL数据库是指各种新型的可扩展／高性能数据库，这类数据库不仅具有 NoSQL数据库对海量数据的存储管理能力，还保持了传统数据库的 ACID和 SQL等特性。

69

在大数据抽取之前，无需清楚数据源的类型和数据的类型，可直接抽取。（）

【A.】√

【B.】 ×

【答案】B

在大数据抽取之前，需要清楚数据源的类型和数据的类型。

70

增量数据抽取机制能够将业务系统中的变化数据按一定的频率准确地捕获到，同时不能对业务系统造成太大的压力，也不能影响现有业务。相对全量数据抽取，增量数据抽取的设计更简单。（）

【A.】√

【B.】 ×

【答案】B

增量数据抽取机制能够将业务系统中的变化数据按一定的频率准确地捕获到，同时不能对业务系统造成太大的压力，也不能影响现有业务。相对全量数据抽取，增量数据抽取的设计更复杂。

71

时间戳是能表示一份数据在某个特定时间之前已经存在的、完整的、可验证的一个数据，其通常是一个字符序列，唯一标识某一刻的时间。（）

【A.】√

【B.】 ×

【答案】A

时间戳是能表示一份数据在某个特定时间之前已经存在的、完整的、可验证的一个数据，其通常是一个字符序列，唯一标识某一刻的时间。

72

使用基于时间戳的增量数据抽取方式进行数据抽取时，系统通过比较上次抽取时间与时间戳字段的值来决定抽取的数据。（）

【A.】√

【B.】 ×

【答案】A

使用基于时间戳的增量数据抽取方式进行数据抽取时，系统通过比较上次抽取时间与时间戳字段的值来决定抽取的数据。

73

时间戳方式的优点是性能优异，系统设计清晰，数据抽取相对复杂，可以实现数据的递增加载。（）

【A.】√

【B.】 ×

【答案】B

时间戳方式的优点是性能优异，系统设计清晰，数据抽取相对简单，可以实现数据的递增加载。

74

数据清洗是数据预处理的重要部分，其主要工作是检查数据的完整性及数据的一致性， 对其中的噪声数据进行平滑，对丢失的数据进行填补，以及对重复的数据进行消除等。（）

【A.】√

【B.】 ×

【答案】A

数据清洗是数据预处理的重要部分，其主要工作是检查数据的完整性及数据的一致性， 对其中的噪声数据进行平滑，对丢失的数据进行填补，以及对重复的数据进行消除等。

75

准确性、完整性、一致性和及时性称为数据质量的四要素。（）

【A.】√

【B.】 ×

【答案】A

我们利用准确性、完整性、一致性和及时性来描述数据质量，通常将其称为数据质量的四要素。

76

数据清洗的目的是消除脏数据，主要消除异常数据、清除重复数据、保证数据的完整性等，进而提高数据的可利用性。（）

【A.】√

【B.】 ×

【答案】A

数据清洗的目的是消除脏数据，主要消除异常数据、清除重复数据、保证数据的完整性等，进而提高数据的可利用性。

77

数据清洗的过程是指通过分析脏数据产生的原因和存在形式，构建数据清洗的模型和算法来完成对脏数据的清除，进而实现将不符合要求的数据转化成满足数据应用要求的数据，为数据分析与建模建立基础。（）

【A.】√

【B.】 ×

【答案】A

数据清洗的过程是指通过分析脏数据产生的原因和存在形式，构建数据清洗的模型和算法来完成对脏数据的清除，进而实现将不符合要求的数据转化成满足数据应用要求的数据，为数据分析与建模建立基础。

78

数据清洗的标准只包含返回率和精确度两个方面。（）

【A.】√

【B.】 ×

【答案】B

数据清洗的衡量标准包括返回率、错误返回率、精确度。

79

k-NN近邻缺失数据填充算法是一种简单快速的算法，它利用本身具有完整记录的属性值实现对缺失属性值的估计。（）

【A.】√

【B.】 ×

【答案】A

k-NN近邻缺失数据填充算法是一种简单快速的算法，它利用本身具有完整记录的属性值实现对缺失属性值的估计。

80

聚类就是将数据集分组为多个类或簇，在同一个簇中的数据对象（记录）之间具有较高的相似度，而不同簇中的对象差别不大。（）

【A.】√

【B.】 ×

【答案】B

聚类就是将数据集分组为多个类或簇，在同一个簇中的数据对象（记录）之间具有较高的相似度，而不同簇中的对象差别就比较大。

81

数据转换可将原始数据转换成适合数据分析的形式，如果数据转换处理不当，将严重扭曲数据本身的内涵，改变数据原本的形态。（）

【A.】√

【B.】 ×

【答案】A

数据转换可将原始数据转换成适合数据分析的形式，如果数据转换处理不当，将严重扭曲数据本身的内涵，改变数据原本的形态。

82

对数转换是将原始数据的自然对数值作为分析数据，如果原始数据中有零，可以在底数中加上一个小数值。（）

【A.】√

【B.】 ×

【答案】A

对数转换是将原始数据的自然对数值作为分析数据，如果原始数据中有零，可以在底数中加上一个小数值。

83

对数转换适用于泊松分布数据。（）

【A.】√

【B.】 ×

【答案】B

对数转换不适用于泊松分布数据。

84

平方根转换适用于泊松分布的数据。（）

【A.】√

【B.】 ×

【答案】A

平方根转换不适用于部分正偏态数据。

85

对数转换适用于轻度偏态数据。（）

【A.】√

【B.】 ×

【答案】B

对数转换不适用于轻度偏态数据。

86

噪声是指测量数据中的随机错误和偏差，通过数据平滑技术可以除去噪声。（）

【A.】√

【B.】 ×

【答案】A

噪声是指测量数据中的随机错误和偏差，通过数据平滑技术可以除去噪声。

87

数据平滑法的处理过程是将获得的实际数据和原始预测数据加权平均，进而去掉数据中的噪声，使得预测结果更接近于真实情况。（）

【A.】√

【B.】 ×

【答案】A

数据平滑法的处理过程是将获得的实际数据和原始预测数据加权平均，进而去掉数据中的噪声，使得预测结果更接近于真实情况。

88

数据平滑法分为移动平均法和分箱平滑法。（）

【A.】√

【B.】 ×

【答案】B

数据平滑法分为移动平均法、指数平滑法和分箱平滑法。

89

移动平均法是预测将来某一时期的平均预测值的一种方法，该方法对过去若干历史数据求算术平均数，并把该数据作为以后的预测值。（）

【A.】√

【B.】 ×

【答案】A

移动平均法是预测将来某一时期的平均预测值的一种方法，该方法对过去若干历史数据求算术平均数，并把该数据作为以后的预测值。

90

移动平均法分为一次移动平均法、混合移动平均法和多次移动平均法。（）

【A.】√

【B.】 ×

【答案】B

移动平均法分为一次移动平均法、二次移动平均法和多次移动平均法。

91

一次移动平均法一般适用于时间序列数据为水平型变动的预测，也适用于明显的长期变动趋势和循环型变动趋势的时间序列预测。（）

【A.】√

【B.】 ×

【答案】B

一次移动平均法一般适用于时间序列数据为水平型变动的预测，而不适用于明显的长期变动趋势和循环型变动趋势的时间序列预测。

92

一次移动平均法仅适用于没有明显的迅速上升或下降趋势的情况,如果时间数列呈直线上升或下降趋势，则需要使用二次移动平均法。（）

【A.】√

【B.】 ×

【答案】A

一次移动平均法仅适用于没有明显的迅速上升或下降趋势的情况,如果时间数列呈直线上升或下降趋势，则需要使用二次移动平均法。

93

二次移动平均法是以历史数据为基础，按时间顺序分段反映后期的变化趋势。（）

【A.】√

【B.】 ×

【答案】A

二次移动平均法是以历史数据为基础，按时间顺序分段反映后期的变化趋势。

94

指数平滑法是预测中常用的方法，这种方法的依据是时间序列的态势具有稳定性或规则性，所以时间序列可顺势推延。（）

【A.】√

【B.】 ×

【答案】A

指数平滑法是预测中常用的方法，这种方法的依据是时间序列的态势具有稳定性或规则性，所以时间序列可顺势推延。

95

指数平滑法适用于中短期发展趋势预测。（）

【A.】√

【B.】 ×

【答案】A

指数平滑法适用于中短期发展趋势预测。

96

分箱平滑法是一种数据局部平滑方法，它是通过考察周围的数据来平滑存储数据，其用箱子的面积来表示不同的箱中的相同个数的数据，用箱的宽度来表示箱中每个数值的取值区间。（）

【A.】√

【B.】 ×

【答案】B

分箱平滑法是一种数据局部平滑方法，它是通过考察周围的数据来平滑存储数据，其用箱子的深度来表示不同的箱中的相同个数的数据，用箱的宽度来表示箱中每个数值的取值区间。

97

分箱平滑法是用箱内数值的平均值、中值或边界值来替代该分箱内各观测的数值。（）

【A.】√

【B.】 ×

【答案】A

分箱平滑法是用箱内数值的平均值、中值或边界值来替代该分箱内各观测的数值。

98

规范化的作用是对重复性的事物和概念，通过规范、规程和制度等达到统一，以获得最佳秩序和效益。（）

【A.】√

【B.】 ×

【答案】A

规范化的作用是对重复性的事物和概念，通过规范、规程和制度等达到统一，以获得最佳秩序和效益。

99

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。（）

【A.】√

【B.】 ×

【答案】A

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。

100

数据规范化方法有最大最小值规范化方法、z分数规范化方法和小数定标规范化方法。（）

【A.】√

【B.】 ×

【答案】A

数据规范化方法有最大最小值规范化方法、z分数规范化方法、小数定标规范化方法。

101

数据约简是指在对挖掘任务和数据本身内容理解的基础之上，寻找依赖于发现目标特征的有用数据，以缩减数据规模，从而在尽可能保持数据原貌的前提下，最大限度地精简数据量。（）

【A.】√

【B.】 ×

【答案】A

数据约简是指在对挖掘任务和数据本身内容理解的基础之上，寻找依赖于发现目标特征的有用数据，以缩减数据规模，从而在尽可能保持数据原貌的前提下，最大限度地精简数据量。

102

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。（）

【A.】√

【B.】 ×

【答案】A

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。

103

确定样本子集大小的因素中不包括计算成本和存储要求。（）

【A.】√

【B.】 ×

【答案】B

确定样本子集大小的因素是计算成本、存储要求、估计量的精度以及与算法和数据特性相关的因素。

104

随机抽样方法的特点是要使总体中每个个体被抽取的可能性都相同。（）

【A.】√

【B.】 ×

【答案】A

随机抽样方法的特点是要使总体中每个个体被抽取的可能性都相同。

105

系统抽样又称为等距抽样，当总体中个体数少，且其分布没有明显的不均匀情况时，常采用系统抽样。（）

【A.】√

【B.】 ×

【答案】B

系统抽样又称为等距抽样，当总体中个体数较多，且其分布没有明显的不均匀情况时，常采用系统抽样。

106

分层抽样又称为类型抽样，是指先将总体单位按主要标志加以分类，分成互不重叠且有限的类型，使其成为层，然后从各层中独立地随机抽取单位。（）

【A.】√

【B.】 ×

【答案】A

分层抽样又称为类型抽样，是指先将总体单位按主要标志加以分类，分成互不重叠且有限的类型，使其成为层，然后从各层中独立地随机抽取单位。

107

数据立方体是二维表格的多维扩展，将三维的数据立方体看作是一组类似的互相叠加起来的二维表格。（）

【A.】√

【B.】 ×

【答案】A

数据立方体是二维表格的多维扩展，将三维的数据立方体看作是一组类似的互相叠加起来的二维表格。

108

在进行高维数据分析时，存在的两个主要困难分别是欧氏距离问题和维数膨胀问题。（）

【A.】√

【B.】 ×

【答案】A

我们在进行高维数据分析时，存在以下两个主要困难。一个问题是欧氏距离问题，一个问题是维数膨胀问题。

109

对于高维数据，通过降维的维数约简方法可以减少冗余数据。（）

【A.】√

【B.】 ×

【答案】A

对于高维数据，通过降维的维数约简方法可以减少冗余数据。

110

维数约简又称为降维，对于较高维空间的数据库Ｘ，通过特征提取或者特征选择的方法，将原空间的维数降至ｍ维。（）

【A.】√

【B.】 ×

【答案】A

维数约简又称为降维，对于较高维空间的数据库Ｘ，通过特征提取或者特征选择的方法，将原空间的维数降至ｍ维。

111

基于数据时序的维数约简可以分为静态维数约简和时间相关维数约简，静态维数约简通常用于处理时间序列。（）

【A.】√

【B.】 ×

【答案】B

基于数据时序的维数约简可以分为静态维数约简和时间相关维数约简，时间相关维数约简通常用于处理时间序列。

112

数值约简是利用替代的方式，使用较小的数据表示替换或估计数据，进而可以减少数据量。（）

【A.】√

【B.】 ×

【答案】A

数值约简是利用替代的方式，使用较小的数据表示替换或估计数据，进而可以减少数据量。

113

数值约简技术分为有参数值约简技术和无参数值约简技术。（）

【A.】√

【B.】 ×

【答案】A

数值约简技术分为有参数值约简技术和无参数值约简技术。

114

无参数值约简技术主要用于存放约简数据的表示，其主要有线性回归模型、直方图、聚类和选择等。（）

【A.】√

【B.】 ×

【答案】B

无参数值约简技术主要用于存放约简数据的表示，其主要有直方图、聚类和选择等。

115

数据集成是应用、存储以及各组织之间传送的数据管理实践活动，其主要考虑合并规整数据问题。（）

【A.】√

【B.】 ×

【答案】A

数据集成是应用、存储以及各组织之间传送的数据管理实践活动，其主要考虑合并规整数据问题。

116

数据集成是指将不同来源、不同格式、不同特点与不同性质的数据在逻辑上或物理上有机地集中，存放在一个一致的数据存储 （如数据仓库）中。（）

【A.】√

【B.】 ×

【答案】A

数据集成是指将不同来源、不同格式、不同特点与不同性质的数据在逻辑上或物理上有机地集中，存放在一个一致的数据存储 （如数据仓库）中。

117

多维可视化的基本方法主要包括基于几何图形、基于图标、基于像素、基于层次结构和基于图结构的混合方法。（）

【A.】√

【B.】 ×

【答案】A

多维可视化的基本方法主要包括基于几何图形、基于图标、基于像素、基于层次结构和基于图结构的混合方法。

118

数据集成最复杂和困难的问题是数据格式转换，也就是将多种数据格式转换为统一的格式。（）

【A.】√

【B.】 ×

【答案】A

数据集成最复杂和困难的问题是数据格式转换，也就是将多种数据格式转换为统一的格式。

119

数据源的异构性一直是困扰数据集成系统的核心问题，异构性的难点主要表现在语法异构和语义异构。（）

【A.】√

【B.】 ×

【答案】A

数据源的异构性一直是困扰数据集成系统的核心问题，异构性的难点主要表现在语法异构和语义异构。

120

持久化数据的安全通过不同层次的管理来实现，即物理层、网络层和数据存储层。（）

【A.】√

【B.】 ×

【答案】B

持久化数据的安全通过不同层次的管理来实现，即物理层、网络层、服务器层、应用层以及数据存储层。

121

大数据分析是指用准确的分析方法和工具来分析经过预处理后的大数据，提取具有价值的信息，进而形成有效的结论，并通过可视化技术展现出来的过程。（）

【A.】√

【B.】 ×

【答案】A

大数据分析是指用准确的分析方法和工具来分析经过预处理后的大数据，提取具有价值的信息，进而形成有效的结论，并通过可视化技术展现出来的过程。

122

大数据挖掘方法以建模理论、数据仓库、机器学习等复合技术为主，数据挖掘是大数据分析的核心，占有重要的地位。（）

【A.】√

【B.】 ×

【答案】A

大数据挖掘方法以建模理论、数据仓库、机器学习等复合技术为主，数据挖掘是大数据分析的核心，占有重要的地位。

123

数据分析的分析目标比较明确，分析条件也比较清楚，基本上就是采用统计方法对数据进行多维度的描述，其从一个假设出发，需要自行选择方程或模型来与假设匹配。而数据挖掘不需要假设，其目标不是很清晰，可以自动建立方程与模型。（）

【A.】√

【B.】 ×

【答案】A

数据分析的分析目标比较明确，分析条件也比较清楚，基本上就是采用统计方法对数据进行多维度的描述，其从一个假设出发，需要自行选择方程或模型来与假设匹配。而数据挖掘不需要假设，其目标不是很清晰，可以自动建立方程与模型。

124

多维数据分析的目标是探索多维数据项的分布规律和模式，并揭示不同维度属性之间的隐含关系。（）

【A.】√

【B.】 ×

【答案】A

多维数据分析的目标是探索多维数据项的分布规律和模式，并揭示不同维度属性之间的隐含关系。

125

从分析的结果上来看，大数据分析主要分为探索性数据分析、证实性数据分析、定性数据分析；从分析的方式上来看，大数据分析主要分为离线数据分析、在线数据分析和交互式分析。（）

【A.】√

【B.】 ×

【答案】A

从分析的结果上来看，大数据分析主要分为探索性数据分析、证实性数据分析、定性数据分析；从分析的方式上来看，大数据分析主要分为离线数据分析、在线数据分析和交互式分析。

126

探索性数据分析是从基于数据本身的角度来说明数据分析方法，涉及模型的假设和统计推断。（）

【A.】√

【B.】 ×

【答案】B

探索性数据分析是从基于数据本身的角度来说明数据分析方法，其并不涉及模型的假设和统计推断。

127

探索性数据分析是从原始数据入手，不完全以实际数据为依据。（）

【A.】√

【B.】 ×

【答案】B

探索性数据分析是从原始数据入手，完全以实际数据为依据。

128

证实性数据分析可以评估观察到的模式或效应的再现性。（）

【A.】√

【B.】 ×

【答案】A

证实性数据分析可以评估观察到的模式或效应的再现性。

129

探索性数据分析强调灵活探求线索和证据，而证实性数据分析则着重评估现有证据。（）

【A.】√

【B.】 ×

【答案】A

探索性数据分析强调灵活探求线索和证据，而证实性数据分析则着重评估现有证据。

130

定性数据分析是指定性研究照片和观察结果等非数值型数据的分析，其是对对象性质特点的一种概括。（）

【A.】√

【B.】 ×

【答案】A

定性数据分析是指定性研究照片和观察结果等非数值型数据的分析，其是对对象性质特点的一种概括。

131

通过数据的数字特征分析，反映数据的集中位置、分散程度、分布形状等，就可以进一步推断出样本中包含的总体信息。（）

【A.】√

【B.】 ×

【答案】A

通过数据的数字特征分析，反映数据的集中位置、分散程度、分布形状等，就可以进一步推断出样本中包含的总体信息。

132

变异系数不能消除单位和平均值不同对两个或多个数据变异程度比较的影响。（）

【A.】√

【B.】 ×

【答案】B

变异系数可以消除单位和平均值不同对两个或多个数据变异程度比较的影响。

133

偏度系数和峰度系数是可刻画数据不对称程度或尾重程度的指标。（）

【A.】√

【B.】 ×

【答案】A

偏度系数和峰度系数是可刻画数据不对称程度或尾重程度的指标。

134

统计学是收集、分析、表述和解释数据的科学，其是指对某一现象数据的搜集、整理、计算、分析、解释和表述等活动。（）

【A.】√

【B.】 ×

【答案】A

统计学是收集、分析、表述和解释数据的科学，其是指对某一现象数据的搜集、整理、计算、分析、解释和表述等活动。

135

模型是指对于某个实际问题或客观事物、规律进行抽象后的一种形式化表达方式。（）

【A.】√

【B.】 ×

【答案】A

模型是所研究的系统、过程、事物或概念的一种表达形式，进一步说，模型是指对于某个实际问题或客观事物、规律进行抽象后的一种形式化表达方式。

136

数据挖掘是大数据分析的核心，其通过建模和构造算法来获取信息与知识。（）

【A.】√

【B.】 ×

【答案】A

数据挖掘是大数据分析的核心，其通过建模和构造算法来获取信息与知识。

137

我们将反映多元线性相关关系的统计指标称为复相关系数。（）

【A.】√

【B.】 ×

【答案】A

将反映多元线性相关关系的统计指标称为复相关系数。

138

回归分析是确定一个随机变量Ｙ对另一个变量Ｘ或一组（Ｘ1，Ｘ２，…，Ｘｋ）变量的相依关系的统计分析方法。（）

【A.】√

【B.】 ×

【答案】A

回归分析是确定一个随机变量Ｙ对另一个变量Ｘ或一组（Ｘ1，Ｘ２，…，Ｘｋ）变量的相依关系的统计分析方法。

139

判别分析是分类方式事先确定，根据若干变量值判断对象归属问题的一种多变量统计分析方法。（）

【A.】√

【B.】 ×

【答案】A

判别分析是分类方式事先确定，根据若干变量值判断对象归属问题的一种多变量统计分析方法。

140

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。（）

【A.】√

【B.】 ×

【答案】A

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。

141

假设分为两种，一种是原假设，另一种是备选假设。（）

【A.】√

【B.】 ×

【答案】A

假设分为两种，一种是原假设，另一种是备选假设。

142

检验假设是数理统计学中根据一定假设条件由样本推断总体的一种方法，先对总体的特征进行某种假设，然后通过抽样研究的统计推理，决定拒绝这个假设还是接受这个假设。（）

【A.】√

【B.】 ×

【答案】A

检验假设是数理统计学中根据一定假设条件由样本推断总体的一种方法，先对总体的特征进行某种假设，然后通过抽样研究的统计推理，决定拒绝这个假设还是接受这个假设。

143

可视化可以使用计算机支持的、交互的方式来表示抽象数据，以增强用户的认知能力，其侧重于通过可视化文本展现数据中隐含的信息和规律，建立符合人的认知规律的心理映像。（）

【A.】√

【B.】 ×

【答案】B

可视化可以使用计算机支持的、交互的方式来表示抽象数据，以增强用户的认知能力，其侧重于通过可视化图形展现数据中隐含的信息和规律，建立符合人的认知规律的心理映像。

144

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

【A.】√

【B.】 ×

【答案】A

由于数据随着时间而变化，可以将数据变化可视化，然后解释导致数据变化的原因。

145

由大及小的可视化展现方式是：先给出一个整体的画面，可以引导阅读者具体深入到一个聚焦的点。（）

【A.】√

【B.】 ×

【答案】A

由大及小的可视化展现方式是：先给出一个整体的画面，可以引导阅读者具体深入到一个聚焦的点。

146

在数据比较的可视化展现中，可以对数据集中突出的不同方面给出一个有力的叙述与说明。（）

【A.】√

【B.】 ×

【答案】A

在数据比较的可视化展现中，可以对数据集中突出的不同方面给出一个有力的叙述与说明。

147

地图就是依据一定的数学法则，使用地图语言、颜色、文字注记等，通过制图综合在一定的载体上，表达地球（或其他天体）上各种事物的空间分布、组合、联系、数量和质量特征及在时间中的发展变化状态绘制的图形，其科学地反映出自然和社会经济现象的分布特征及其相互关系。（）

【A.】√

【B.】 ×

【答案】A

地图就是依据一定的数学法则，使用地图语言、颜色、文字注记等，通过制图综合在一定的载体上，表达地球（或其他天体）上各种事物的空间分布、组合、联系、数量和质量特征及在时间中的发展变化状态绘制的图形，其科学地反映出自然和社会经济现象的分布特征及其相互关系。

148

地图的构成要素只包括图形要素和数学要素。（）

【A.】√

【B.】 ×

【答案】B

地图的构成要素主要包括图形要素、数学要素和辅助要素等。

149

数学要素用来确定地学要素的空间相关位置，是地图内容骨架的要素。（）

【A.】√

【B.】 ×

【答案】A

数学要素用来确定地学要素的空间相关位置，是地图内容骨架的要素。

150

地图是地理信息的载体，其容纳和储存了巨大数量的信息，而作为信息的载体，只能是传统概念上的纸质地图、实体模型。（）

【A.】√

【B.】 ×

【答案】B

地图是地理信息的载体，其容纳和储存了巨大数量的信息，而作为信息的载体，可以是传统概念上的纸质地图、实体模型，可以是各种可视化屏幕影像、声像地图，也可以是触觉地图。

151

将抽象的指标数据转换成容易感知的数据时，用户便更容易理解图形要表达的意义。（）

【A.】√

【B.】 ×

【答案】A

将抽象的指标数据转换成容易感知的数据时，用户便更容易理解图形要表达的意义。

152

气泡图是柱状图的一种变体，通过每个点的面积大小来反映第三维。（）

【A.】√

【B.】 ×

【答案】B

气泡图是散点图的一种变体，通过每个点的面积大小来反映第三维。

153

根据信息的特征可以将信息可视化分为一维信息可视化、二维信息可视化、三维信息可视化、多维信息可视化、层次信息可视化、网络信息可视化、时序信息可视化。（）

【A.】√

【B.】 ×

【答案】A

根据信息的特征可以将信息可视化分为一维信息可视化、二维信息可视化、三维信息可视化、多维信息可视化、层次信息可视化、网络信息可视化、时序信息可视化。

154

文本数据是典型的结构化数据类型。（）

【A.】√

【B.】 ×

【答案】B

文本数据是典型的非结构化数据类型。

155

文本可视化可以将文本中蕴含的语义特征直观地展示出来，这些语义特征主要有词频与重要度、逻辑结构、主题聚类、动态演化规律等。（）

【A.】√

【B.】 ×

【答案】A

文本可视化可以将文本中蕴含的语义特征直观地展示出来，这些语义特征主要有词频与重要度、逻辑结构、主题聚类、动态演化规律等。

156

时空数据是带有地理位置与时间标签的数据。（）

【A.】√

【B.】 ×

【答案】A

时空数据是带有地理位置与时间标签的数据。

157

维数灾难通常是指在涉及向量计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象。（）

【A.】√

【B.】 ×

【答案】A

维数灾难通常是指在涉及向量计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象。

158

平方根转换适用于等比数据。（）

【A.】√

【B.】 ×

【答案】B

平方根转换不适用于等比数据。

三、问答题

1

Cyber空间与信息化是什么？

【答案】Cyber空间本意是指以计算机技术、现代通信网络技术、虚拟现实技术等信息技术的综合运用为基础，以知识和信息为内容的新型空间，是一个人工世界。信息化是将现实世界中的事物和现象以数据的形式存储到 Cyber空间中，是一个数据生产的过程。

2

数据科学的定义？

【答案】数据科学是关于数据的科学或者研究数据的科学，用来研究 Cyber空间中数据界奥秘的理论、方法和技术，其研究的对象是数据界中的数据。主要包括两个方面：一方面是研究数据本身，研究数据的各种类型、状态、属性及变化形式和变化规律；另一方面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的现象和规律。

3

数据加工的定义？

【答案】为了提升数据质量、降低数据计算的复杂度、减少数据计算量以及提升数据处理的精准度，数据科学项目需要对原始数据进行一定的加工处理工作，主要包括数据清洗、数据变换、数据集成、数据脱敏、数据约简和数据标注等。

4

互联网大数据的定义是？

【答案】来自互联网的网络大数据是指 “人、机、物”三元世界在网络空间中交互、融合所产生并在互联网上可获得的大数据。

5

结构化数据、非结构化数据的定义及区别是？

【答案】结构化数据是具有数据结构描述信息的数据；非结构化数据是不方便使用固定结构来表现的数据。区别：结构化数据先有结构，再有数据；非结构化数据只有数据，没有结构。

6

大数据的处理过程是？

【答案】大数据处理的全过程可以概括为５个步骤，分别是大数据获取与存储管理，大数据抽取与清洗，大数据约简与集成，大数据分析与挖掘，大数据分析结果解释与可视化展现。

7

请简述MapReduce的计算过程？

【答案】MapReduce由Map和Reduce两个阶段组成。Map函数以键值对 （key，value）作为输入，产生另外一系列键值对作为中间输出写入本地硬盘。MapReduce框架会自动将这些中间数据按照 key 值进行聚集，将 key值相同的数据统一交给 Reduce函数处理。Reduce函数以 key及对应的 value列表作为输入，经合并 key相同的 value值后，产生另外一系列键值对作为最终输出写 入 HDFS。

8

在大数据的可视分析中，用户界面与交互设计需要考虑的因素有哪些？

【答案】用户驱动的数据简化；可扩展性与多级层次；表示证据和不确定性；异构数据融合；交互查询中的数据概要与分流；时变特征分析；设计与工程开发。

9

根据信息的特征，可以将信息可视化划分为哪些类别？

【答案】根据信息的特征可以将信息可视化分为一维信息可视化、二维信息可视化、三维信息可视化、多维信息可视化、层次信息可视化、网络信息可视化、时序信息可视化。

10

请简述MapReduce的基本原理。

【答案】将大数据分成小块逐个分析，最后将提取出来的数据汇总分析，进而获得需要的结果。

11

MapReduce模型的编程方法步骤？

【答案】（１）遍历输入数据，并将之解析成键值对 （key，value）。（２）将输入键值对 （key，value）映射 （Map）成另外一些键值对 （key，value）。（３）依据 key对中间数据进行分组。（４）以组为单位对数据进行约简 （Reduce）。（５）将最终产生的键值对 （key，value）保存到输出文件中。

12

地图的构成要素有哪些？请简要说明并解释每种要素的含义。

【答案】地图的构成要素主要包括图形要素、数学要素和辅助要素等。图形要素：图形要素是地图根据制图的要求所表达的内容，包括注记、地学基础。数学要素：数学要素用来确定地学要素的空间相关位置，是地图内容骨架的要素。辅助要素：辅助要素说明地图编制状况及为方便地图应用所必须提供的内容。

13

大数据获取的定义？

【答案】大数据获取又称为大数据采集，其是利用数据获取工具，从系统外部获取数据，并存入系统内部的存储资源。

14

传统数据获取与大数据获取的区别？

【答案】传统数据获取：数据来源单一，数据量相对较小，结构单一；大数据获取：数据来源广泛，数据量巨大，包括结构化数据、半结构化数据和非结构化数据。

15

图像数字化的定义？

【答案】图像数字化是将模拟图像转换为数字图像，数字图像便于计算机进行存储与处理。图像数字化是进行数字图像处理的前提，其必须以图像的电子化作为基础，把模拟图像转变成电子信号，随后才将其转换成数字图像信号。

16

大数据获取的挑战主要包括哪些方面？

【答案】数据源多种多样；数据量巨大；数据变化快；保证数据获取的可靠性；避免重复数据；保证数据的真实性。

17

什么是网站数据采集？

【答案】网站数据采集是指通过网络采集软件工具或网站公开 API（Application Programming Interface，应用程序编程接口）等方式将网站上的非结构化数据、半结构化数据和结构化数据从网页中提取出来，并将其存储到统一的本地数据文件中。采集的数据包括图片、音频、视频等。

18

通用网络爬虫的局限性？

【答案】不同领域、不同背景的用户具有不同的检索目的和需求，而通用网络爬虫所返回的结果可能含有大量用户并不需要的网页。通用网络爬虫的目标是获得尽可能大的网络覆盖率，从而造成了有限的网络爬虫服务器资源与无限的网络数据资源之间的冲突。图片、数据库、音频、视频多媒体等不同类型的非结构化数据大量出现，通用网络爬虫对这些信息含量密集数据的获取出现了困难。通用网络爬虫主要提供基于关键字的检索，难以支持基于语义信息的查询。

19

NewSQL系统的主要架构？

【答案】一种架构是数据库工作在一个分布式集群的节点上，其中每个节点拥有一个数据子集，其将 SQL查询分成查询片段发送给自己所在的数据的节点上执行，可以通过添加节点来线性扩展。另一种架构是数据库系统有一个单一的主节点的数据源，有一组节点用来做事务处理，这些节点接到特定的 SQL 查询后，将把它所需的所有数据从主节点上取回来后执行 SQL查询，再返回结果。

20

大数据抽取的定义？

【答案】大数据抽取过程是搜索全部数据源，按照某种标准选择合乎要求的数据，并将被选中的数据传送到目的地中存储。简单地说，大数据抽取过程就是从数据源中抽取数据并传送到目的地数据系统中的过程。

21

构建大数据抽取程序需要考虑哪些条件？

【答案】抽取数据对象的类型；在数据源中寻找所需的数据对象的方法；为已找到的数据选择组装格式；将找到的数据对象组装到数据库中的方法；生成和维护大数据抽取过程所需的元数据；一般不单独设计组装算法。

22

常用的大数据的抽取方法有哪些？

【答案】同构同质数据抽取；同构异质数据抽取；文件型数据抽取；全量数据抽取；增量数据抽取。

23

同构异质数据库的定义是？

【答案】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。

24

增量数据抽取的特点是？

【答案】只抽取发生变化的数据；相对于全量数据抽取更为快捷，处理量更少；采用增量数据抽取需要与数据装载时的更新策略相对应。

25

请简述数据平滑法的处理过程。

【答案】数据平滑法的处理过程是将获得的实际数据和原始预测数据加权平均，进而去掉数据中的噪声，使得预测结果更接近于真实情况，数据平滑法是趋势法或时间序列法的一种具体应用，数据平滑法分为移动平均法、指数平滑法和分箱平滑法。

26

移动平均法的定义是？

【答案】移动平均法是预测将来某一时期的平均预测值的一种方法。该方法对过去若干历史数据求算术平均数，并把该数据作为以后的预测值。移动平均法分为一次移动平均法、二次移动平均法和多次移动平均法。

27

指数平滑法中平滑常数的选取规则是？

【答案】当时间序列呈现较稳定的水平趋势时，应选较小的α值，一般可在 0.05~0.20 取值；当时间序列有波动，但长期趋势变化不大时，可选稍大的α值，一般可在 0.1~0.4取值；当时间序列波动很大，长期趋势变化幅度较大，呈现明显且迅速的上升或下降趋势时，宜选择较大的α值，如可在 0.6~0.8取值，以使预测模型灵敏度高些，能迅速跟上数据的变化；当时间序列数据是上升或下降的趋势时，α应在 0.6~1取较大的值。

28

数据规范化的定义和作用是？

【答案】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。规范化的作用是对重复性的事物和概念，通过规范、规程和制度等达到统一，以获得最佳秩序和效益。

29

分箱平滑法的定义是？

【答案】分箱平滑法是一种数据局部平滑方法，它是通过考察周围的数据来平滑存储数据。其用箱子的深度来表示不同的箱中的相同个数的数据，用箱的宽度来表示箱中每个数值的取值区间。

30

z分数规范化方法的步骤是？

【答案】求出各变量的算术平均值 （数学期望）ｘ和标准差 ｓ；依据均值和标准差对原始数据进行标准化；将逆指标前的正负号对调。

31

数据约简的定义是？

【答案】数据约简是指在对挖掘任务和数据本身内容理解的基础之上，寻找依赖于发现目标特征的有用数据，以缩减数据规模，从而在尽可能保持数据原貌的前提下，最大限度地精简数据量。

32

请简述随机抽样方法的特点和实现过程。

【答案】随机抽样方法的特点是要使总体中每个个体被抽取的可能性都相同。当总体中的个体数较少时，常采用抽签的方法抽取样本，即将总体中的每个个体依次编上号码 １，２，３，…， ｍ，制作一套与总体中每个个体号码相对应的、形状大小相同的卡片号签，并将卡片号签均匀搅拌，从中抽出一个卡片号签，这个卡片号签所对应的每个个体就组成一个样本。

33

请简述系统抽样的特点和实现过程。

【答案】系统抽样又称为等距抽样，当总体中个体数较多，且其分布没有明显的不均匀情况时，常采用系统抽样。一般情况下，可将总体分成均衡的若干部分，然后按照预先定好的规则，从每一部分抽取相同个数的个体，这样的抽样即系统抽样。

34

分层抽样的定义是？

【答案】分层抽样又称为类型抽样，是指先将总体单位按主要标志加以分类，分成互不重叠且有限的类型，使其成为层，然后从各层中独立地随机抽取单位。

35

什么是数值约简？数值约简技术可以分为哪两类？

【答案】数值约简是利用替代的方式，使用较小的数据表示替换或估计数据，进而可以减少数据量。数值约简技术分为有参数值约简技术和无参数值约简技术。有参数值约简技术是使用模型来评估数据，其只使用参数，而不是实际值。无参数值约简技术主要用于存放约简数据的表示。

36

请简述数据集成的概念。

【答案】数据集成是应用、存储以及各组织之间传送的数据管理实践活动，其主要考虑合并规整数据问题。 数据集成是指将不同来源、不同格式、不同特点与不同性质的数据在逻辑上或物理上有机地集中，存放在一个一致的数据存储（如数据仓库）中。

37

大数据分析的定义是？

【答案】大数据分析是指用准确的分析方法和工具来分析经过预处理后的大数据，提取具有价值的信息，进而形成有效的结论，并通过可视化技术展现出来的过程。

38

数据分析与数据挖掘的主要区别是？

【答案】数据分析通常是分析以往的数据，或者评价某时间段内取得的效果。而数据挖掘的数据量极大，要依靠挖掘算法来找出隐藏在大量数据中的规律和模式，也就是从数据中提取出隐含的、未知的、有价值的信息；数据分析的分析目标比较明确，分析条件也比较清楚，而数据挖掘不需要假设，其目标不是很清晰，可以自动建立方程与模型；数据分析通常针对数字化的数据，而数据挖掘可以采用不同类型的数据，如声音和文本等；数据分析是针对历史数据分析得出各项指标，为决策提供数据支持，而数据挖掘是对数据分析加机器决策，为将来的事件提供决策；数据分析对结果进行解释，呈现出有效信息。数据挖掘的结果不容易解释，对信息进行价值评估，着眼于预测未来，并提出决策性建议；数据分析是把数据变成信息的工具，数据挖掘是把信息变成认知的工具，如果需要从数据中获取一定的规律，需要数据分析和数据挖掘结合使用。

39

什么是模型？模型由哪三个部分组成？

【答案】模型是所研究的系统、过程、事物或概念的一种表达形式，进一步说，模型是指对于某个实际问题或客观事物、规律进行抽象后的一种形式化表达方式。任何模型都是由目标、变量和关系三个部分组成。

40

什么是相关分析？

【答案】相关分析是研究现象之间是否存在某种依存关系，并对有依存关系的现象，探讨其相关方向以及相关程度，其是研究概率变量之间的相关性的一种统计方法。

41

请简述回归分析的步骤。

【答案】确定自变量与因变量。根据自变量与因变量的历史统计数据进行计算，建立回归分析预测模型。获得自变量与因变量之间的某种因果关系。模型检验，预测误差，小误差表明模型可以得到比较好的预测结果。运用已确定的回归预测模型进行预测计算，再根据具体的实际数据，运用相关知识进行全面分析，进而得到最终的预测值。

42

相关分析与回归分析的基本区别是？

【答案】相关分析研究的是现象之间是否相关，以及相关的方向和密切程度，不区分是自变量或因变量。而回归分析则要分析现象之间相关的具体形式，并用数学模型来表现其具体因果关系。

43

请简述检验假设的方法。

【答案】根据问题的需要对所研究的总体作某种假设Ｈ。选取合适的统计量，这个统计量的选取要使得在假设Ｈ成立时，其分布为已知。由实测的样本计算出统计量的值。根据预先给定的显著性水平进行检验，作出拒绝或接受假设Ｈ的判断。

44

什么是检验假设？

【答案】检验假设是数理统计学中根据一定假设条件由样本推断总体的一种方法，先对总体的特征进行某种假设，然后通过抽样研究的统计推理，决定拒绝这个假设还是接受这个假设。

45

可视化的定义是？

【答案】可视化是利用计算机图形学和图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并利用数据分析和开发工具发现其中未知信息的交互处理的理论、方法和技术。