林子雨 编著

《大数据导论》

教材配套习题和答案

2020 年 4 月

第 1 章　大数据概述

一、单选题

1、 下面关于数据的说法，错误的是：（B）

A. 数据的根本价值在于可以为人们找出答案
B. 数据的价值会因为不断使用而削减
C. 数据的价值会因为不断重组而产生更大的价值
D. 目前阶段，数据的产生不以人的意志为转移

2、 第 3 次信息化浪潮的标志是：（C）

A. 个人计算机的普及
B. 互联网的普及
C. 云计算、大数据和物联网技术的普及
D. 人工智能的普及

3、 物联网的发展最终导致了人类社会数据量的第三次跃升，使得数据产生方式进入了：（D）

A. 手工创建阶段
B. 运营式系统阶段
C. 用户原创内容阶段
D. 感知式系统阶段

4、 英国的大数据发展战略是：(D)

A. 稳步实施“三步走”战略，打造面向未来的大数据创新生态
B. 通过发展创新性解决方案并应用于实践来促进大数据发展
C. 以大数据等技术为核心应对第四次工业革命
D. 紧抓大数据产业机遇，应对脱欧后的经济挑战

5. 以下哪个不是大数据的“4V”特性：（D）

A. 数据量大
B. 数据类型繁多
C. 处理速度快
D. 价值密度高

二、多选题

1、 数据的类型主要包括：（ABCD）

A. 文本

B. 图片
C. 音频
D. 视频

2、计算机系统中的数据组织形式主要有两种，分别是：（AD）

A. 文件
B. 视频
C. 音频
D. 数据库

3、为了让数据变得可用，需要对数据进行三个步骤的处理，分别是：（ACD）

A. 数据清洗
B. 数据抽样
C. 数据管理
D. 数据分析

4、信息科技为大数据时代提供技术支撑，主要体现在哪三个方面：（ABD）

A. 存储设备容量不断增加
B. CPU 处理能力大幅提升
C. 量子计算机全面普及
D. 网络带宽不断增加

5、人类社会的数据产生方式大致经历了哪三个阶段：（BCD）

A. 手工生产阶段
B. 运营式系统阶段
C. 用户原创内容阶段
D. 感知式系统阶段

6、关于“大数据摩尔定律”，以下说法正确的是：（ABC）

A. 人类社会产生的数据一直都在以每年 50%的速度增长
B. 人类社会的数据量大约每两年就增加一倍
C. 人类在最近两年产生的数据量相当于之前产生的全部数据量之和
D. 人类社会的数据量以每年 10%的速度增长

7、人类自古以来在科学研究上先后历经了哪几种范式：（ABCD）

A．实验科学
B．理论科学
C．计算科学
D．数据密集型科学

8、大数据将会对社会发展产生深远的影响，具体表现在以下哪几个方面：（ABCD）

A．大数据决策成为一种新的决策方式
B．大数据成为提升国家治理能力的新途径
C．大数据应用促进信息技术与各行业的深度融合
D．大数据开发推动新技术和新应用的不断涌现

9、大数据产业是指一切与支撑大数据组织管理和价值发现相关的企业经济活动的集合。以下哪些属于大数据产业的某个环节（ABCD）：

A．IT 基础设施层
B．数据源层
C．数据管理层
D．数据分析层

第 2 章　大数据与其他新兴技术之间的关系

一、单选题

1、早期的云计算产品 AWS 是由哪家企业提出的：（C）

A．IBM
B．微软
C．亚马逊
D．谷歌

2、云计算包括 3 种类型。面向所有用户提供服务，只要是注册付费的用户都可以使用，这种云计算属于：（A）

A．公有云
B．私有云
C．混合云
D．独立云

3、云计算包括 3 种类型。只为特定用户提供服务，比如大型企业出于安全考虑自建的云环境，只为企业内部提供服务，这种云计算属于：（B）

A．公有云

B．私有云
C．混合云
D．独立云

4、以下关于大数据、云计算和物联网的区别，描述错误的是：（C）

A．大数据侧重于对海量数据的存储、处理与分析，从海量数据中发现价值，服务于生产和生活
B．云计算本质上旨在整合和优化各种 IT 资源并通过网络以服务的方式，廉价地提供给用户
C．云计算旨在从海量数据中发现价值，服务于生产和生活
D．物联网的发展目标是实现物物相连，应用创新是物联网发展的核心

5、以下关于机器学习，描述错误的是：(C)

A．是一门涉及统计学、系统辨识、逼近理论、神经网络、优化理论、计算机科学、脑科学等诸多领域的交叉学科
B．研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能
C．机器学习强调三个关键词：算法、模型、训练
D．基于数据的机器学习是现代智能技术中的重要方法之一

6、以下关于知识图谱，描述错误的是：（C）

A．又称为科学知识图谱
B．在图书情报界称为知识域可视化或知识领域映射地图
C．知识图谱属于密码学研究范畴
D．知识图谱可用于反欺诈、不一致性验证、组团欺诈等公共安全保障领域

7、以下关于人机交互，描述错误的是：（B）

A．人机交互是一门研究系统与用户之间的交互关系的学科
B．人机交互界面通常是指用户不可见的部分
C．系统可以是各种各样的机器，也可以是计算机化的系统和软件
D．用户通过人机交互界面与系统交流，并进行操作

8、以下关于计算机视觉，描述错误的是：（D）

A．计算机视觉是一门研究如何使机器“看”的科学
B．是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量的机器视觉
C．计算机视觉是一门综合性的学科
D．语音识别属于计算机视觉的典型应用

9、关于大数据与区块链的联系，下面描述错误的是：（D）
A．区块链使大数据极大降低信用成本
B．区块链是构建大数据时代的信任基石
C．区块链是促进大数据价值流通的管道
D．区块链会提升大数据的信用成本

二、多选题

1、 传统的 IT 资源获取方式的主要缺点是：（ABC）

A． 初期成本高，建设周期长
B． 后期需要自己维护，使用成本高
C． IT 资源供应量有限
D． IT 资源供应量无限

2、 云计算的主要优点是：（BCD）

A． 初期投入大，需要用户自己维护
B． 初期零成本，瞬时可获得
C． 后期免维护，使用成本低
D． 在供应 IT 资源量方面“予取予求”

3、 云计算包括哪 3 种典型的服务模式：（ABD）

A． IaaS（基础设施即服务）
B． PaaS（平台即服务）
C． MaaS（机器即服务）
D． SaaS（软件即服务）

4、 云计算包括哪 3 种类型：（ACD）

A． 公有云
B． 独立云
C． 私有云
D． 混合云

5、从技术架构上看，物联网主要包括哪几层：（ABCD）

A． 感知层
B． 网络层
C． 处理层
D． 应用层

5、 以下关于大数据、云计算和物联网的联系，描述正确的是：（ABCD）

A．从整体上看，大数据、云计算和物联网这三者是相辅相成的
B．大数据根植于云计算，大数据分析的很多技术都来自于云计算
C．大数据为云计算提供了“用武之地”
D．物联网需要借助于云计算和大数据技术，实现物联网大数据的存储、分析和处理

6、 以下关于大数据与人工智能的联系，描述正确的是：（ABCD）

A．人工智能需要数据来建立其智能，特别是机器学习
B．人工智能应用的数据越多，其获得的结果就越准确
C．大数据为人工智能提供了海量的数据，使得人工智能技术有了长足的发展
D．大数据技术为人工智能提供了强大的存储能力和计算能力

7、下面关于比特币和区块链之间关系的描述，正确的是：（BC）
A．比特币和区块链没有任何关系
B．区块链是比特币的底层技术
C．比特币是区块链的一种应用
D．比特币是比区块链更先进的一种技术

8、比特币要解决的两个核心问题是：（AD）
A.防篡改
B．防丢失
C．防贬值
D．去中心化记账

9、在比特币区块链中关于如何争夺记账权的问题，下面描述正确的是：（ABCD）
A．采用的是 POW 机制，也就是“工作量证明机制”
B．记账节点通过计算数学题，来争夺记账权
C．对于数学公式的计算，除了从零开始遍历随机数碰运气以外，没有其他办法
D．解题的过程，又叫“挖矿”，记账节点被称为矿工。谁先解对，谁就获得记账权

10、区块链的三要素是：（ABC）
A．交易
B．区块
C．链
D．比特币

第 3 章 大数据基础知识

一、单选题

1、下面关于大数据安全问题，描述错误的是：（D）

A. 大数据的价值并不单纯地来源于它的用途，而更多地源自其二次利用
B. 对大数据的收集、处理、保存不当，会加剧数据信息泄露的风险
C. 大数据成为国家之间博弈的新战场
D. 大数据对于国家安全没有产生影响

2、下面关于棱镜门事件描述错误的是：（C）

A. 棱镜计划（PRISM） 是一项由美国国家安全局（NSA）自 2007 年起开始实施的绝密电子监听计划
B. 在该计划中，美国国家安全局和联邦调查局利用平台和技术上的优势，开展全球范围内的监听活动
C. 该计划的目的是为了促进世界和平与发展
D. 该计划对全世界重点地区、部门、公司甚至个人进行布控

3、下面关于手机软件采集个人信息的描述错误的是：（C）

A. 在我们的日常生活中，部分手机 APP 往往会“私自窃密”
B. 有的 APP 在提供服务时，采取特殊方式来获得用户授权，这本质上仍属“未经同意”
C. 在微信朋友圈广泛传播的各种测试小程序是安全的，不会窃取用户个人信息
D. 手机 APP 过度采集个人信息呈现普遍趋势，最突出的是在非必要的情况下获取位置信息和访问联系人权限

4、下面描述错误的是：（D）

A. “探针盒子”就是一款自动收集用户隐私的产品
B. 许多顾客在使用 WiFi 之后会收到大量的广告信息，甚至自己的手机号码也会被当做信息进行多次买卖
C. 在免费上网的背后，其实也存在着不小的信息安全风险，或许一不小心，就落入了电脑黑客们设计的 WiFi 陷阱之中
D. 免费 WIFI 都是安全的，可以放心使用

5、下面关于机械思维的核心思想，描述错误的是：（B）

A. 世界变化的规律是确定的
B. 世界变化的规律是无法确定的
C. 规律不仅是可以被认识的，而且可以用简单的公式或者语言描述清楚
D. 这些规律应该是放之四海而皆准的，可以应用到各种未知领域指导实践

6、我们在使用智能手机进行导航来避开城市拥堵路段时，体现了哪种大数据思维方式：（A）

A．我为人人，人人为我
B．全样而非抽样
C．效率而非精确
D．相关而非因果

7、谷歌采用搜索引擎大数据进行流感趋势预测，体现了哪种大数据思维方式：（B）

A．我为人人，人人为我
B．全样而非抽样
C．效率而非精确
D．相关而非因果

8、“啤酒与尿布”的故事，体现了哪种大数据思维方式：（D）

A．我为人人，人人为我
B．全样而非抽样
C．效率而非精确
D．相关而非因果

9、大数据的简单算法比小数据的复杂算法更有效，体现了哪种大数据思维方式：（A）

A．以数据为中心
B．全样而非抽样
C．效率而非精确
D．相关而非因果

10、迪士尼 MagicBand 手环，体现了哪种大数据思维方式：（A）

A．我为人人，人人为我
B．全样而非抽样
C．效率而非精确
相关而非因果

11、下面关于大数据伦理的描述，错误的是：（D）

A．大数据伦理属于科技伦理的范畴
B．大数据伦理问题是指由于大数据技术的产生和使用而引发的社会问题
C．作为一种新的技术，大数据技术像其他所有技术一样，其本身是无所谓好坏的，而它的“善”与“恶”全然在于对大数据技术的使用者

D．大数据技术本身就存在“善”和“恶”的区分

12、现在的互联网，基于大数据和人工智能的推荐应用越来越多，越来越深入，我们一直被“喂食着”经过智能化筛选推荐的信息，久而久之，会导致什么问题：（A）

A、信息茧房问题
B、隐形偏差问题
C、大数据杀熟问题
D、隐私泄露问题

13、　　下面哪一个不属于大数据伦理问题：（D）

A．隐私泄露问题
B．数据安全问题
C．数字鸿沟问题
D．数据冗余问题

14、下面关于政府数据孤岛描述错误的是：（D）

A. 有些政府部门错误地将数据资源等同于一般资源，认为占有就是财富，热衷于搜集，但不愿共享
B. 有些部门只盯着自己的数据服务系统，结果因为数据标准、系统接口等技术原因，无法与外单位、外部门联通
C. 有些地方，对大数据缺乏顶层设计，导致各条线、各部门固有的本位主义作祟，壁垒林立，数据无法流动
D. 即使涉及到工作机密、商业机密，政府也应该毫不保留地共享数据

15、关于推进数据共享开放的描述，错误的是：（D）

A．要改变政府职能部门“数据孤岛”现象，立足于数据资源的共享互换，设定相对明确的数据标准，实现部门之间的数据对接与共享
B．要使不同省区市之间的数据实现对接与共享，解决数据“画地为牢”的问题，实现数据共享共用
C．在企业内部，破除“数据孤岛”，推进数据融合
D．不同企业之间，为了保护各自商业利益，不宜实现数据共享

16、下面关于数据权的描述，错误的是：（D）

A．数据权的概念发起于英国，主要将其视为信息社会的一项基本公民权利

B．数据权包括两个方面：数据主权和数据权利
C．数据主权的主体是国家，是一个国家独立自主对本国数据进行管理和利用的权力
D．数据主权的主体是公民，是相对应于公民数据采集义务而形成的对数据利用的权利

17、下面关于政府信息公开与政府数据开放的描述，错误的是：（B）

A．政府信息公开与政府数据开放是一对既相互区别又相互联系的概念
B．信息是没有经过任何加工与解读的原始记录，没有明确的含义，而数据则是经过加工处理并被赋予一定含义的
C．政府信息公开主要是为了对公众知情权的满足而出现的
D．政府数据开放强调的是数据的再利用，公众可以分享数据利用创造的经济和社会价值

18、关于公民的隐私权，下面描述错误的是：（A）

A．修改权是隐私权利人具有的依法了解自身信息资料是否被行政主体利用的权利
B．支配权是隐私权利人的基本权利之一，隐私权利人对自己的个人信息的收集、储存、传播、使用、开放等享有支配权
C．保障权是指公民有权要求政府在数据开放的过程中保障涉及其个人隐私的信息资料不被开放、不被滥用和不被泄露
D．救济权是公民在自身的合法权益受到侵害时，按照法定程序采取法律手段维护自身权益的权利

19、关于大数据交易在发展过程中遇到的问题，下面描述错误的是：（D）

A．互联网数据马太效应显现
B．市场信用体系缺失、监管有待加强
C．大数据交易规则和标准缺乏
D．数据质量评价与估值定价已经很完善

20、目前大数据交易市场上存在很多种定价机制，但是不包括以下哪项：（D）

A．平台预定价
B．自动计价
C．拍卖式定价
D．随机性定价

21、我国首家大数据交易所是：（A）

A．贵阳大数据交易所
B．上海数据交易中心
C．华东江苏大数据交易中心

D．浙江大数据交易中心

二、多选题

1、传统的数据安全的威胁主要包括：（ABC）

A．计算机病毒
B．黑客攻击
C．数据信息存储介质的损坏
D．数据复制

2、大数据安全表现出与传统数据安全不同的特征，具体来说包括哪几个方面：（ABCD）

A．大数据成为网络攻击的显著目标
B．大数据加大隐私泄露风险
C．大数据技术被应用到攻击手段中
D．大数据成为高级可持续攻击（APT）的载体

3、舍恩伯格在《大数据时代：生活、工作与思维的大变革》一书中明确指出，大数据时代最大的转变就是思维方式的 3 种转变，具体包括：（ABC）

A．全样而非抽样
B．效率而非精确
C．相关而非因果
D．务实而非务虚

4、下面关于搜索引擎“点击模型”的描述正确的是：（ABCD）

A．随着数据量的积累，点击模型对搜索结果排名的预测越来越准确，它的重要性也越来越大
B．点击模型的准确性取决于数据量的大小
C．一个搜索引擎使用的时间越长，数据的积累就越充分，对于长尾搜索就做得越准确
当整个搜索行业都意识到点击数据的重要性后，这个市场上的竞争就从技术竞争变成了数据竞争

5、下面关于隐私泄露问题的描述，正确的是：（ABCD）

A．大数据时代下的隐私与传统隐私的最大区别在于隐私的数据化，即隐私主要以“个人数据”的形式出现
B．用户在使用搜索引擎时，搜索引擎可以精确地刻画出该用户的“数字肖像”
C．通过数据预测，可以预测个体“未来的隐私”
D．“数据痕迹”往往永远无法彻底消除，会被永久保留记录

6、下面关于数字鸿沟问题的描述，正确的是：（ACD）

A．数字鸿沟被认为是信息时代的“马太效应”，即先进技术的成果不能为人公正分享，于是造成“富者越富、穷者越穷”的情况
B．数字鸿沟因为大数据技术的诞生而趋向弥合
C．数字鸿沟是一个涉及公平公正的问题
D．在我国，东中西部地区、城乡之间等都可以明显感受到数字鸿沟的存在

7、下面关于数据独裁的描述，正确的是：（ABCD）

A．所谓的“数据独裁”是指在大数据时代，由于数据量的爆炸式增长，导致做出判断和选择的难度徒增，迫使人们必须完全依赖数据的预测和结论才能做出最终的决策
B．从某个角度来讲，数据独裁就是让数据统治人类，使人类彻底走向唯数据主义
C．数据独裁最终将导致人类思维被“空心化”，进而是创新意识的丧失
D．数据独裁还可能使人们丧失了人的自主意识、反思和批判的能力，最终沦为数据的奴隶

8、因数据而产生的垄断问题，主要包括哪几种类型：（ABCD）

A．数据可能造成进入壁垒或扩张壁垒
B．拥有大数据形成市场支配地位并滥用
C．因数据产品而形成市场支配地位并滥用
D．涉及数据方面的垄断协议

9、企业数据孤岛产生的原因主要包括哪两个方面：（AB）

A．以功能为标准的部门划分导致数据孤岛
B．不同类型、不同版本的信息化管理系统导致数据孤岛
C．机构设置不合理
D．各个部门责权利不清晰

10、消除数据孤岛对于政府具有哪些重要的意义：（ABCD）

A．有助于提升资源利用率
B．有助于推动政府转型
C．有助于提高行政效率
D．有助于促进跨部门合作

11、消除数据孤岛对于企业具有哪些重要的意义：（ABC）

A．有助于企业做出有利于生产要素组合优化的决策，使企业能够合理配置资源，实现企业利益最大化

B．有利于企业获得更好的经营发展能力
C．企业信息的增多可以增加做出正确选择的能力，从而提高经济效率
D．不利于企业长远的发展

12、实现数据共享，在政府层面面临的挑战包括：（ABCD）

A．不愿共享开放
B．不敢共享开放
C．不会共享开放
D．数据中心共享开放作用不强

13、　　实现数据共享，在企业层面面临的挑战包括：（ABC）

A．系统孤岛挑战
B．组织架构挑战
C．数据合作挑战
D．利润风险挑战

14、关于政府数据开放的意义，下面描述正确的是：（ABC）

A．政府开放数据有利于促进开放透明政府的形成
B．政府开放数据有利于创新创业和经济增长
C．政府开放数据有利于社会治理创新
D．政府开放数据将会对政府正常运作产生威胁

15、目前进行数据交易的形式主要包括哪几种：（ABC）

A．大数据交易公司
B．数据交易所
C．API 模式
D．PPT 模式

16、大数据交易平台的类型主要包括哪两种：（AD）

A．综合数据服务平台
B．实时数据交易平时
C．零散数据交易平台
D．第三方数据交易平台

17、交易数据的来源主要包括哪些：（ABCD）

A．政府公开数据

B．企业内部数据
C．数据供应方数据
D．网页爬虫数据

18、交易产品的类型主要包括哪几种：（ABCD）

A．API
B．数据包
C．云服务
D．解决方案

19、大数据交易平台的运营模式主要包括哪两种：（BC）

A．具有交易实时显示功能的交易平台
B．兼具中介和数据处理加工功能的交易平台
C．只具备中介功能的交易平台
D．只具备数据处理加工功能的交易平台

20、可以从哪些维度评价数据价值：（ABCD）

A．数据样本量
B．数据品种
C．数据完整性
D．数据实时性

第 4 章 大数据应用

一、单选题

1、下面关于推荐系统的描述错误的是：（D）

A．推荐系统是自动联系用户和物品的一种工具
B．和搜索引擎相比，推荐系统通过研究用户的兴趣偏好，进行个性化计算
C．推荐系统可发现用户的兴趣点，帮助用户从海量信息中去发掘自己潜在的需求
D．推荐系统是一种只能通过专家进行人工推荐的系统

2、以下推荐方法中，哪一个是基于内容的推荐：（C）

A．由资深的专业人士来进行物品的筛选和推荐
B．基于统计信息进行推荐
C．通过机器学习的方法去描述内容的特征，并基于内容的特征来发现与之相似的内容
D．对多种推荐算法进行有机组合，然后给出推荐结果

3、以下哪项不属于大数据在城市管理中的应用：（D）

A．智能交通
B．环保监测
C．城市规划
D．比赛预测

4、以下哪项不属于大数据在零售领域的应用：（A）

A．大数据征信
B．发现关联购物行为
C．客户群体划分
D．供应链管理

二、多选题

1、一个完整的推荐系统通常包括哪 3 个组成模块：（ABC）

A．用户建模模块
B．推荐对象建模模块
C．推荐算法模块
D．可视化模块

2、智慧医疗具有哪些优点：（ABCD）

A．促进优质医疗资源的共享
B．避免患者重复检查
C．促进医疗智能化
D．有助于实现全民免费医疗

3、下面关于智能物流的描述，正确的是：ABCD

A．又称智慧物流，是利用智能化技术，使物流系统能模仿人的智能，具有思维、感知、学习、推理判断和自行解决物流中某些问题的能力
B．可以帮助实现物流资源优化调度和有效配置，并且提升物流系统效率
C．智能物流概念源自 2010 年 IBM 发布的研究报告《智慧的未来供应链》
D．智能物流概念经历了自动化、信息化、网络化 3 个发展阶段

4、智能物流具有哪几个方面的重要作用：（ABC）

A．提高物流的信息化和智能化水平
B．降低物流成本和提高物流效率
C．提高物流活动的一体化
D．提高了物流的复杂性

5、大数据在金融领域的应用主要包括：（ABCD）

A．高频交易
B．市场情绪分析
C．信贷风险分析
D．大数据征信

6、大数据在餐饮行业的应用主要包括：（ABCD）

A．大数据驱动的团购模式
B．利用大数据为用户推荐消费内容
C．利用大数据调整线下门店布局
D．利用大数据控制店内人流量

第 5 章 数据采集与预处理

一、单选题

1、 以下哪个步骤不属于数据的采集与预处理：（D）

A．利用 ETL 工具将分布的、异构数据源中的数据，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中
B．利用日志采集工具把实时采集的数据作为流计算系统的输入，进行实时处理分析
C．利用网页爬虫程序到互联网网站中爬取数据
D、对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据

2、 以下哪项不属于数据清洗的内容：（B）

A．一致性检查
B．精确度校验
C．无效值和缺失值的处理
D．成对删除

3、以下哪个不是 Flume 的核心组件：（A）
A．数据块（Block）
B．数据源（Source）
C．数据通道（Channel）
D．数据槽（Sink）

6、下面关于网络爬虫的描述错误的是：（D）
A．网络爬虫是一个自动提取网页的程序

B．为搜索引擎从万维网上下载网页，是搜索引擎的重要组成部分
C．爬虫从一个或若干个初始网页的 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，直到满足系统的一定停止条件
D．网络爬虫的行为和人们访问网站的行为是完全不同的

7、下面关于网络爬虫的描述正确的是：（D）
A．网络爬虫由控制节点、爬虫节点和资源库构成
B．网络爬虫中可以有多个控制节点，每个控制节点下可以有多个爬虫节点
C．控制节点之间可以互相通信，控制节点和其下的各爬虫节点之间也可以进行互相通信
D．属于同一个控制节点下的各爬虫节点间不可以互相通信

8、以下哪个不是 Scrapy 体系架构的组成部分：（C）
A．Scrapy 引擎（Engine）
B. 爬虫（Spiders）
C. 支持者（Support）
D. 下载器（Downloader）

9、下面关于反爬机制描述错误的是：（D）
A．简单低级的网络爬虫，数据采集速度快，伪装度低，如果没有反爬机制，它们可以很快地抓取大量数据，甚至因为请求过多，造成网站服务器不能正常工作，影响了企业的业务开展
B．反爬机制也是一把双刃剑，一方面可以保护企业网站和网站数据，但是，另一方面，如果反爬机制过于严格，可能会误伤到真正的用户请求
C．如果既要和“网络爬虫”死磕，又要保证很低的误伤率，那么又会增加网站研发的成本
D．反爬机制不利于信息的自由流通，不利于网站发展，应该坚决取消

10、假设有一个数据集 X={4,8,15,21,21,24,25,28,34}，这里采用基于平均值的等高分箱方法对其进行平滑处理，则分箱处理结果是：（B）
A. {8,8,8,22,22,22,29,29,29}
B. {9,9,9,22,22,22,29,29,29}
C. {9,9,9,21,21,21,29,29,29}
D. {9,9,9,22,22,22,28,28,28}

11、假设属性的最大值和最小值分别是 87000 元和 11000 元，现在需要利用 Min-Max 规范化方法，将“顾客收入”属性的值映射到 0~1 的范围内，则“顾客收入”属性的值为 72400 元时，对应的转换结果是：（A）
A．0.808
B．0.837
C．0.769
D．0.987

12、假设 A 班级的平均分是 80，标准差是 10，A 考了 90 分；B 班的平均分是 400，标准差是 100，B 考了 600 分。采用 Z-Score 规范化以后，二者谁的成绩更加优秀：（B）
A．A 的成绩更为优秀

B. B 的成绩更为优秀
C.二者一样优秀
D．无法比较

13、假设属性的取值范围是-957~924，当属性的值为 426 时，采用小数定标规范化方法对应的转换结果是：（C）
A．0.421
B.0.433
C. 0.426
D.0.489

二、多选题
1、数据采集的三大要点是：（ABC）

A． 全面性
B． 多维性
C． 高效性
D． 精确性

2、数据采集的主要数据源包括：（ABCD）

A． 传感器数据
B． 互联网数据
C． 日志文件
D． 企业业务系统数据

3、需要清洗的数据的主要类型包括：（ACD）
A． 残缺数据
B． 干净数据
C． 错误数据
D． 重复数据

4、典型的数据采集方法包括：（ABCD）
A． 系统日志采集
B．分布式消息订阅分发
C．ETL
D．网络数据采集

5、Kafka 的架构包括哪些组件：（ABCD）
A. 话题（Topic）
B. 生产者（Producer）

C. 服务代理（Broker）
D. 消费者（Consumer）
6、网络爬虫的类型主要包括：（）
A．通用网络爬虫
B. 聚焦网络爬虫
C. 增量式网络爬虫
D. 深层网络爬虫

7、常见的数据转换策略包括：（ABCD）
A．平滑处理
B．聚集处理
C．数据泛化处理
D．规范化处理

8、常用的规范化处理方法包括：（ABD）
A． Min-Max 规范化
B.Z-Score 规范化
C.曲面规范化
D.小数定标规范化

9、数据脱敏的主要原则包括：（ABCD）
A. 保持原有数据特征
B. 保持数据之间的一致性
C. 保持业务规则的关联性
D. 多次脱敏之间的数据一致性

10、数据脱敏的方法主要包括：（ABCD）
A. 数据替换
B. 无效化
C. 随机化
D. 偏移和取整

第 6 章 数据存储与管理

一、单选题
1、以下哪项不属于传统的数据存储和管理技术：（A）

A． NoSQL 数据库
B． 文件系统
C． 关系数据库
D． 数据仓库

2、以下关于分布式文件系统，描述错误的是：（B）

A．是一种通过网络实现文件在多台主机上进行分布式存储的文件系统
B．所有的分布式文件系统的设计都是采用“客户机/服务器”（Client/Server）模式
C．谷歌开发了分布式文件系统 GFS
D．Hadoop 分布式文件系统（Hadoop Distributed File System，HDFS）是针对 GFS 的开源实现

3、以下描述错误的是：（D）

A．传统的关系数据库可以较好地支持结构化数据存储和管理
B．Web 2.0 的迅猛发展以及大数据时代的到来，使关系数据库的发展越来越力不从心
C．传统的关系数据库由于数据模型不灵活、水平扩展能力较差等局限性，已经无法满足各种类型的非结构化数据的大规模存储需求
D．传统关系数据库引以为豪的一些关键特性，如事务机制和支持复杂查询，在 Web 2.0 时代成为不可或缺的核心特性

4、以下关于 NoSQL 数据库描述错误的是：（C）

A、NoSQL 是一种不同于关系数据库的数据库管理系统设计方式，是对非关系型数据库的统称
B、NoSQL 所采用的数据模型并非传统关系数据库的关系模型，而是类似键/值、列族、文档等非关系模型
C、NoSQL 数据库有固定的表结构，通常存在较多连接操作
D、与关系数据库相比，NoSQL 具有灵活的水平可扩展性，可以支持海量数据存储

5、在数据库的发展历史上，先后出现过多种数据库类型，但是，不包括：（B）
A.网状数据库
B.球形数据库
C.层次数据库
D.关系数据库

6、下面关于关系数据库特点的描述，错误的是：（D）
A．采用表格的储存方式，数据以行和列的方式进行存储，要读取和查询都十分方便
B. 为了规范化数据、减少重复数据以及充分利用好存储空间，把数据按照最小关系表的形式进行存储
C. 由于关系数据库将数据存储在数据表中，数据操作的瓶颈出现在多张数据表的操作中，而且数据表越多这个问题越严重
D. 关系数据库采用非结构化查询语言来对数据库进行查询

7、下面关于 NewSQL 数据库的描述，错误的是：（B）
A. NewSQL 数据库保持了传统数据库支持 ACID 和 SQL 等特性
B. 不同的 NewSQL 数据库的内部结构基本相同

C. 都支持关系数据模型
D.都使用 SQL 作为其主要的接口

8、下面关于 Hadoop 的描述错误的是：（C）
A. Hadoop 是一个能够对大量数据进行分布式处理的软件框架
B. 作为并行分布式计算平台，Hadoop 采用分布式存储和分布式处理两大核心技术，能够高效地处理 PB 级数据
C.Hadoop 只支持 Java 编程语言
D. Hadoop 可以高效稳定地运行在廉价的计算机集群上，可以扩展到数以千计的计算机节点上

9、下面哪个不是 Hadoop 生态系统的组件：（B）
A.HDFS
B.SQL Server
C.MapReduce
D.HBase

10、下面组件哪个是负责在 Hadoop 和关系数据库之间实现数据导入导出的：（C）
A.MySQL
B.HDFS
C.Sqoop
D.Flume

11、下面组件哪个是负责分布式资源调度与管理的：（A）
A.YARN
B.Flume
C.Zookeeper
D.Kafka

12、下面组件哪个是数据挖掘库：（B）
A.Zookeeper
B.Mahout
C.MySQL
D.HBase

13、下面组件哪个是负责日志收集的：（D）
A.Ambari
B.Zookeeper
C.HDFS
D.Flume

14、下面组件哪个是负责 Hadoop 集群的安装、部署、配置和管理的：（C）
A. Kafka

B.YARN
C.Ambari
D.Flume

15、下列哪一项不属于 NoSQL 的四大类型：（D）
A.文档数据库
B.图数据库
C.列族数据库
D.时间戳数据库

16、下列关于键值数据库的描述，哪一项是错误的：（D）
A.扩展性好，灵活性好
B.大量写操作时性能高
C.无法存储结构化信息
D.条件查询效率高

17、下列关于列族数据库的描述，哪一项是错误的：（A）
A.查找速度慢，可扩展性差
B.功能较少，大都不支持强事务一致性
C.容易进行分布式扩展
D.复杂性低

18、关于文档数据库的说法，下列哪一项是错误的：（A）
A 数据是规则的
B.性能好（高并发）
C.缺乏统一的查询语法
D.复杂性低

19、下列关于云数据库的描述，哪个是错误的？(C)
A.云数据库是部署和虚拟化在云计算环境中的数据库
B.云数据库是在云计算的大背景下发展起来的一种新兴的共享基础架构的方法
C.云数据库价格不菲，维护费用极其昂贵
D.云数据库具有高可扩展性、高可用性、采用多租形式和支持资源有效分发等特点

20、下列哪一个不属于云数据库产品？(A)
A.本地安装 MySQL
B.阿里云 RDS
C.Oracle Cloud
D.百度云数据库

21、下面哪一项不是云数据库的特性？(B)

A.动态可扩展
B.高成本
C.易用性
D.大规模并行处理

22、下列关于 BigTable 的描述，哪个是错误的？(A)
A.爬虫持续不断地抓取新页面，这些页面每隔一段时间地存储到 BigTable 里
B.BigTable 是一个分布式存储系统
C.BigTable 起初用于解决典型的互联网搜索问题
D.网络搜索应用查询建立好的索引，从 BigTable 得到网页

二、多选题

1、数据仓库的特性包括：（ABCD）

A． 面向主题的
B． 集成的
C． 相对稳定的
D． 反映历史变化的

2、NoSQL 数据库具有以下几个特点：（ABC）

A． 灵活的可扩展性
B． 灵活的数据模型
C． 与云计算紧密融合
D． 数据模型比较死板

3、一个典型的数据仓库系统通常包含哪几个组成部分：（ABCD）
A. 数据源
B. 数据存储和管理
C. OLAP 服务器
D. 前端工具和应用

4、下面关于并行数据库的描述正确的是：（ABD）
A. 并行数据库是指那些在无共享的体系结构中进行数据操作的数据库系统
B. 大部分采用了关系数据模型并且支持 SQL 语句查询
C. 并行数据库系统具有较好的弹性
D. 并行数据库的另一个问题就是系统的容错性较差

5、Hadoop 的特性主要包括：（ABC）
A. 高可靠性
B. 高可扩展性
C. 高容错性
D.成本高

6、HDFS 要实现哪些设计目标：（BCD）
A.复杂的文件模型
B. 兼容廉价的硬件设备
C. 流数据读写
D. 强大的跨平台兼容性

7、HDFS 的局限性包括：（ACD）
A. 不适合低延迟数据访问
B.无法用于大规模数据存储
C. 无法高效存储大量小文件
D. 不支持多用户写入及任意修改文件

8、下面关于 HDFS 的体系结构描述正确的是：（ABC）
A. HDFS 采用了主从（Master/Slave）结构模型，一个 HDFS 集群包括一个名称节点和若干个数据节点
B. 名称节点作为中心服务器，负责管理文件系统的命名空间及客户端对文件的访问
C. 集群中的数据节点一般是一个节点运行一个数据节点进程，负责处理文件系统客户端的读/写请求
D. 名称节点会周期性地向数据节点发送“心跳”信息，报告自己的状态

9、 下列关于文档数据库的描述，哪些是正确的？（AD）
A.性能好（高并发），灵活性高
B.具备统一的查询语法
C.文档数据库支持文档间的事务
D.复杂性低，数据结构灵活

10、下列关于图数据库的描述，哪些是正确的？(ABCD)
A.专门用于处理具有高度相互关联关系的数据
B.比较适合于社交网络、模式识别、依赖分析、推荐系统以及路径寻找等问题
C.灵活性高，支持复杂的图算法
D.复杂性高，只能支持一定的数据规模

11、下列关于数据模型的描述，哪些是正确的？(ABCD)
A.HBase 采用表来组织数据，表由行和列组成，列划分为若干个列族
B.每个 HBase 表都由若干行组成，每个行由行键（row key）来标识
C.列族里的数据通过列限定符（或列）来定位

D.每个单元格都保存着同一份数据的多个版本，这些版本采用时间戳进行索引

12、HBase 的系统架构包括哪几个组成部分：(ABCD)
A.客户端
B.Zookeeper 服务器
C.Master 主服务器
D.Region 服务器

13、下面关于 Google Spanner 的描述正确的是：（ABCD）
A. Spanner 是一个可扩展的、全球分布式的数据库
B. 在最高抽象层面，Spanner 就是一个数据库，把数据分片存储在许多 Paxos 状态机上，这些机器位于遍布全球的数据中心内
C. 随着数据的变化和服务器的变化，Spanner 会自动把数据进行重新分片，从而有效应对负载变化和处理失败
D. Spanner 被设计成可以扩展到几百万个机器节点，跨越成百上千个数据中心，具备几万亿数据库行的规模

第 7 章 数据处理与分析

一、单选题
1、下面描述错误的是：（C）
A. 数据分析可以分为广义的数据分析和狭义的数据分析
B.广义的数据分析就包括狭义的数据分析和数据挖掘。
C.数据挖掘就是指狭义的数据分析
D. 数据挖掘是指从大量的数据中挖掘出未知的、且有价值的信息和知识的过程

2、下面描述错误的是：（A）
A. 数据挖掘的目标明确，先做假设，然后通过数据分析来验证假设是否正确，从而得到相应的结论
B. 数据挖掘的重点在寻找未知的模式与规律
C. 数据分析一般都是得到一个指标统计量结果，如总和、平均值等
D. 数据挖掘则是输出模型或规则，并且可相应得到模型得分或标签

3、下面关于机器学习和数据挖掘的描述错误的是：（D）
A. 机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科
B. 数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程。
C. 数据挖掘可以视为机器学习与数据库的交叉
D.数据挖掘是机器学习的底层技术

4、以下哪个不是典型的分类方法：（C）

A.决策树
B.朴素贝叶斯
C. K-Means
D.人工神经网络

5、以下哪个不是聚类方法：（D）
A. GMM
B. LDA
C. DBSCAN
D.TPLINK

6、聚类分析的常见应用场景不包括：（A）
A. 发现关联购买行为
B. 目标用户的群体分类
C. 不同产品的价值组合
D. 探测发现离群点和异常值

7、下面关于回归分析的描述错误的是：（C）
A. 是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法
B. 回归分析按照涉及的变量的多少，分为一元回归和多元回归分析
C. 按照因变量的多少，可分为线性回归分析和非线性回归分析
D. 在大数据分析中，回归分析是一种预测性的建模技术

8、下面关于协同过滤算法的描述错误的是：（D）
A. 基于用户的协同过滤算法（简称 UserCF 算法）是推荐系统中最古老的算法，可以说，UserCF 的诞生标志着推荐系统的诞生
B. 基于物品的协同过滤算法（简称 ItemCF 算法）是目前业界应用最多的算法
C. 基于模型的协同过滤算法（ModelCF）是通过已经观察到的所有用户给产品的打分，来推断每个用户的喜好并向用户推荐适合的产品
D. UserCF 算法是给目标用户推荐那些和他们之前喜欢的物品相似的物品。

9、下面属于批处理技术的是：(A)

A． MapReduce
B． Storm
C． Spark Streaming
D． GraphX

10、下面属于流计算技术的是：(C)

A. Spark MLLib
B. GraphX
C. S4

D. Hive

11、下面属于图计算技术的是：（A）

A. Pregel
B. Dremel
C. Impala
D. DStream

12、下面属于查询分析计算技术的是：（C）

A. Spark Streaming
B. Storm
C. Hive
D. Pregel

13、下列关于流计算的说法，哪项是错误的？(D)
A.实时获取来自不同数据源的海量数据，经过实时分析处理，获得有价值的信息
B.流计算秉承一个基本理念，即 数据的价值随着时间的流逝而降低
C.对于一个流计算系统来说，它应该支持 TB 级甚至是 PB 级的数据规模
D.流计算只需要保证较低的延迟时间，即只达到秒级别即可处理一切问题

14.下列关于数据处理流程，说法有误的是？(D)
A.在传统的数据处理流程中，存储的数据是旧的
B.在传统的数据处理流程中，需要用户主动发出查询来获取结果
C.传统的数据处理流程，需要先采集数据并存储在关系数据库等数据管理系统中
D.流计算的处理流程一般包含三个阶段：数据实时采集、数据批量计算、实时查询服务

15、下面哪个属于图数据库：（A）
A. Neo4j
B.MySQL
C.HBase
D.Oracle

16、下列关于 MapReduce 模型的描述，错误的是哪一项？(D)
A.MapReduce 采用“ 分而治之”策略
B.MapReduce 设计的一个理念就是“ 计算向数据靠拢”
C.MapReduce 框架采用了 Master/Slave 架构
D.MapReduce 应用程序只能用 Java 来写

17、关于数据仓库 Impala 的描述错误的是：（D）
A. Impala 作为开源大数据分析引擎，支持实时计算，它提供了与 Hive 类似的功能，并在性

能上比 Hive 高出 3~30 倍
B. Impala 是由 Cloudera 公司开发的查询系统
C. Impala 提供了 SQL 语义，能查询存储在 Hadoop 的 HDFS 和 HBase 上的 PB 级别海量数据
D. Impala 最初是参照 MySQL 系统进行设计的

18、下面关于 Spark 和 Hadoop 的关系，描述错误的是：（D）
A. Spark 和 Hadoop 一样，既包含了存储的组件，也包含了计算的组件
B. Spark 作为计算框架，只能解决数据计算问题，无法解决数据存储问题
C. Spark 只是取代了 Hadoop 生态系统中的计算框架 MapReduce，而 Hadoop 中的其他组件依然在企业大数据系统中发挥着重要的作用
D. 越来越多的企业放弃 MapReduce，转而使用 Spark 开发企业应用

19、以下哪个不是 Spark 的生态系统的组件：（C）
A. Spark Streaming
B. Structured Streaming
C. Zookeeper
D. GraphX

20、以下哪个组件是 Spark 中的机器学习算法库：（A）
A. MLlib
B. Spark Core
C.Machine Leaning
D.Spark SQL

21、以下哪个组件是 Spark 中用于结构化数据处理的组件：（A）
A.Spark SQL
B.Spark Core
C.Spark Streaming
D.Structured Streaming

22、Shark 与 Spark SQL 的关系是：（B）
A.二者没有任何关系
B.Shark 是 Spark SQL 的前身
C.Spark SQL 是 Shark 的前身
D.二者是一个软件的两个不同名称，本质上是一个东西

23、下面关于 TensorFlow 和 TensorFlowOnSpark 的描述错误的是：（B）
A. TensorFlow 是一个采用数据流图（Data Flow Graph）、用于数值计算的开源软件库
B. TensorFlow 是一个开源的、基于 Java 的机器学习框架
C. TensorFlowOnSpark 项目是由 Yahoo 开源的一个软件包，能将 TensorFlow 与 Spark 结合在一起使用
D. TensorFlowOnSpark 为 Apache Hadoop 和 Apache Spark 集群带来可扩展的深度学习功能

24、以下哪个不是 Storm 的特点：（D）
A. 可扩展性
B. 可靠的消息处理
C. 支持各种编程语言
D. 复杂的 API

25、下面关于 Spark Streaming 和 Storm 的描述错误的是：（A）
A. Spark Streaming 可以实现毫秒级的流计算
B. Storm 可以实现毫秒级响应
C. Spark Streaming 构建在 Spark Core 之上
D. Spark Streaming 可以同时兼容批量和实时数据处理的逻辑和算法

26、下面关于 Flink 的描述错误的是：（C）
A. Flink 是一个针对流数据和批数据的分布式计算框架
B. Flink 的设计思想主要来源于 Hadoop、MPP 数据库、流计算系统等
C. Flink 主要是由 Python 代码实现的
D. Flink 所要处理的主要场景是流数据，批数据只是流数据的一个特例而已

二、多选题

1、数据分析主要实现哪三大作用：（BCD）
A. 误差分析
B. 现状分析
C. 原因分析
D. 预测分析

2、数据挖掘主要侧重解决哪几类问题：（ABCD）
A. 分类
B. 聚类
C. 关联
D. 预测

3、下面关于数据分析与数据处理的描述，正确的是：（ACD）
A. 数据分析过程通常会伴随着发生数据处理（或者说伴随着大量数据计算）
B. 数据分析和数据处理不存在紧密的关联关系
C. 二者是融合在一起的，很难割裂开来
D. 当用户在进行数据分析的时候，底层的计算机系统会根据数据分析任务的要求，使用程序进行大量的数据处理

4、下面关于大数据处理与分析的描述，正确的是：（ABCD）
A. 在理论层面，数据分析需要统计学、机器学习和数据挖掘等知识
B. 在技术层面，包括单机分析工具（比如 SPSS、SAS 等）或单机编程语言（比如 Python、R），

以及大数据处理与分析技术（比如 MapReduce、Spark、Hive 等）
C. 在大数据时代到来之前，数据分析主要以小规模的抽样数据为主，一般使用单机分析工具（比如 SPSS 和 SAS）或者单机编程（比如 Python、R）的方式来实现分析程序
D. 到了大数据时代，数据量爆炸式地增长，数据分析就需要采用分布式实现技术，比如使用 MapReduce、Spark 或 Flink 编写分布式分析程序，借助于集群的多台机器进行并行数据处理分析

5、常见的关联规则挖掘算法包括：(BC)
A. MP-Growth 算法
B.FP-Growth 算法
C. Apriori 算法
D.Bpriori 算法

6、协同过滤主要包括：（ABC）
A.基于用户的协同过滤
B.基于物品的协同过滤
C.基于模型的协同过滤
D.基于分类的协同过滤

7、大数据处理分析技术主要包括哪几种类型：（ABCD）
A. 批处理计算
B. 流计算
C. 图计算
D. 查询分析计算

8、一次 BSP 计算过程包括一系列全局超步（超步就是指计算中的一次迭代），每个超步主要包括哪几个组件：（ACD）
A. 局部计算
B. 中间计算
C. 通信
D. 栅栏同步

9、下面关于 MapReduce 工作流程的描述，正确的是：（ABD）
A. 一个大的 MapReduce 作业，会被拆分成许多个 Map 任务在多台机器上并行执行
B. 每个 Map 任务通常运行在数据存储的节点上
C. 当 Map 任务结束后，会生成以<key,value-list>形式表示的许多中间结果
D. Reduce 任务会对中间结果进行汇总计算得到最后结果

10、Hadoop 的 MapReduce 的缺点包括：（ABC）
A. 表达能力有限
B. 磁盘 IO 开销大

C. 延迟高
D. 中间结果多

11、Hive 底层所依赖的计算引擎可以是：（BCD）
A.Flink
B.MapReduce
C.Tez
D.Spark

12、下面关于 Hive 的描述正确的是：(ABCD)
A. Hive 是一个基于 Hadoop 的数据仓库工具，可以用于对存储在 Hadoop 文件中的数据集进行数据整理、特殊查询和分析处理
B. Hive 的学习门槛比较低，因为它提供了类似于关系数据库 SQL 语言的查询语言——HiveQL
C. 当采用 MapReduce 作为执行引擎时，Hive 可以通过 HiveQL 语句快速实现简单的 MapReduce 统计，Hive 自身可以将 HiveQL 语句快速转换成 MapReduce 任务进行运行
D. Hive 在某种程度上可以看作是用户编程接口，其本身并不存储和处理数据

13、关于 Hive 与 Hadoop 生态系统中其他组件的关系，下面描述正确的是：（ABC）
A. HDFS 作为高可靠的底层存储，用来存储海量数据
B. MapReduce 对这些海量数据进行批处理，实现高性能计算
C. 用 HiveQL 语句编写的处理逻辑，最终都要转化为 MapReduce 任务来运行
D. Hive 的目标是取代 HBase

14、Hive 的系统架构主要包括哪几个模块：(BCD)
A. 探查模块
B. 驱动模块
C. 元数据存储模块
D. 用户接口模块

15、关于数据仓库 Impala 的描述正确的是：（BC）
A. Impala 是由 Oracle 公司开发的查询系统
B. 与 Hive 类似，Impala 也可以直接与 HDFS 和 HBase 进行交互
C. Impala 采用了与商用 MPP 并行关系数据库类似的分布式查询引擎，可以直接从 HDFS 或者 HBase 中用 SQL 语句查询数据，而不需要把 SQL 语句转化成 MapReduce 任务来执行
D. Impala 和 Hive 采用了不同的 SQL 语法、ODBC 驱动程序和用户接口

16、Spark 的特点主要包括：（ABC）
A. 运行速度快
B. 容易使用
C. 通用性
D. 运行模式单一

17、Spark 相对于 MapReduce 的优点包括：（ABD）

A. Spark 的计算模式也属于 MapReduce，但不局限于 Map 和 Reduce 操作，还提供了多种数据集操作类型，编程模型比 MapReduce 更灵活
B. Spark 提供了内存计算，中间结果直接放到内存中，带来了更高的迭代运算效率
C.Spark 同时提供了存储功能，而 MapReduce 不支持存储
D. Spark 基于 DAG 的任务调度执行机制，要优于 MapReduce 的迭代执行机制

18、不同的计算框架统一运行在 YARN 中，可以带来哪些好处：（BCD）
A. 减少了所使用的编程语言的种类
B. 计算资源按需伸缩
C. 不用负载应用混搭，集群利用率高
D. 共享底层存储，避免数据跨集群迁移

19、在实际应用中，大数据处理主要包括哪几种类型：（ABC）
A. 复杂的批量数据处理：时间跨度通常在数十分钟到数小时之间
B. 基于历史数据的交互式查询：时间跨度通常在数十秒到数分钟之间
C. 基于实时数据流的数据处理：时间跨度通常在数百毫秒到数秒之间
D.基于历史数据的流查询：时间跨度在数十秒到数分钟之间

20、下面关于 Spark 的运行架构的描述，正确的是：（ABD）
A. Spark 运行架构包括 Cluster Manager、Worker Node、Driver Program 和 Executor
B.Spark 集群资源管理器可以是 Spark 自带的资源管理器，也可以是 YARN 或 Mesos 等资源管理框架
C. Spark 采用“P2P 架构”
D. Spark 利用多线程来执行具体的任务

21 下面关于 RDD 的描述正确的是：（ABC）
A. 一个 RDD 就是一个分布式对象集合
B. 一个 RDD 本质上是一个只读的分区记录集合
C. RDD 提供了一组丰富的操作以支持常见的数据运算，分为“行动”（Action）和“转换”（Transformation）两种类型
D. RDD 不适合对于数据集中元素执行相同操作的批处理式应用，而比较适合用于需要异步、细粒度状态的应用

22、Spark 的集群部署方式包括：（ABC）
A. Spark on Mesos 模式
B. Spark on YARN 模式
C. Spark on Kubernetes 模式
D. Local 模式

23、下面关于 Spark SQL 的描述正确的是：（ACD）
A. Spark SQL 在 Hive 兼容层面仅依赖 HiveQL 解析和 Hive 元数据
B. Spark SQL 目前支持 Scala、Java 编程语言，暂时不支持 Python 语言
C. Spark SQL 执行计划生成和优化都由 Catalyst（函数式关系查询优化框架）负责
D. Spark SQL 增加了 DataFrame（即带有 Schema 信息的 RDD），使用户可以在 Spark SQL

中执行 SQL 语句

24、下面关于 Spark Streaming 的描述正确的是：（ABCD）
A. Spark Streaming 是构建在 Spark Core 上的实时计算框架，它扩展了 Spark 处理大规模流式数据的能力
B. Spark Streaming 可结合批处理和交互查询，适合一些需要对历史数据和实时数据进行结合分析的应用场景
C. Spark Streaming 可整合多种输入数据源，如 Kafka、Flume、HDFS，甚至是普通的 TCP 套接字
D. Spark Streaming 实际上是以一系列微小批处理来模拟流计算

25、Structured Streaming 包括哪两种处理模型：（AD）
A. 微批处理
B. 高阶处理
C. 分层处理
D. 持续处理

26、关于 Structured Streaming、Spark SQL、Spark Streaming，下面描述正确的是：（ACD）
A. Structured Streaming 处理的数据跟 Spark Streaming 一样，也是源源不断的数据流
B. Spark Streaming 采用的数据抽象是 DataFrame，Structured Streaming 采用的数据抽象是 DStream
C. Structured Streaming 可以使用 Spark SQL 的 DataFrame/Dataset 来处理数据流
D. Spark SQL 只能处理静态的数据，而 Structured Streaming 可以处理结构化的数据流

27、Spark MLlib 主要提供了哪几个方面的工具：（ABCD）
A. 算法工具
B. 特征化工具
C. 流水线
D. 实用工具

28、下面关于 Storm 框架设计描述正确的是：（ABD）
A. Storm 运行在分布式集群中，其运行任务的方式与 Hadoop 类似
B. 在 Hadoop 上运行的是 MapReduce 作业，而在 Storm 上运行的是“Topology”
C. Storm 集群采用 P2P 架构
D. Storm 采用了 Zookeeper 来作为分布式协调组件

29、下面关于 Flink 的描述正确的是：（BCD）
A. Flink 和 Spark 一样，都是基于磁盘的计算框架
B. 当全部运行在 Hadoop YARN 之上时，Flink 的性能甚至还要略好于 Spark
C. Flink 的流计算性能和 Storm 差不多，可以支持毫秒级的响应
D. Spark 的市场影响力和社区活跃度明显超过 Flink

30、Flink 系统主要由哪两个组件组成：（AB）

A. JobManager
B. TaskManager
C.JobTracker
D.TaskTracker

31、下面关于大数据编程框架 Beam 的描述正确的是：（BCD）
A. Beam 是由微软公司贡献的 Apache 顶级项目
B. Beam 的目标是为开发者提供一个易于使用、却又很强大的数据并行处理模型，能够支持流处理和批处理
C. Beam 是一个开源的统一的编程模型，开发者可以使用 Beam SDK 来创建数据处理管道，然后，这些程序可以在任何支持的执行引擎上运行
D. Beam SDK 定义了开发分布式数据处理任务业务逻辑的 API 接口，即提供一个统一的编程接口给到上层应用的开发者

32、查询分析系统 Dremel 的特点主要包括：（BD）
A. Dremel 是一个面向小规模数据的、稳定的系统
B. Dremel 的数据模型是嵌套的
C. Dremel 中的数据是用行式存储的
D. Dremel 结合了 Web 搜索和并行 DBMS 的技术

第 8 章 数据可视化

一、单选题

1、 下列关于数据可视化的描述，哪个是错误的？(D)
A.数据可视化是指将大型数据集中的数据以图形图像形式表示
B.利用数据分析和开发工具发现其中未知信息的处理过程
C.数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示
D.将数据的各个属性值以一维数据的形式表示
2、 下列哪个不属于可视化工具？(D)
A.Google Chart API
B.D3
C.Visual.ly
D.Spark
3、 下列说法错误的是？(B)
A.大数据魔镜是一款优秀的国产数据分析软件，可以让用户真正理解探索分析数据
B.Tableau 是桌面系统中最简单的商业智能工具软件，是一个用于网页作图、生成互动图形的 JavaScript 函数库
C.Google Fusion Tables 让一般使用者也可以轻松制作出专业的统计地图
D.Modest Maps 是一个小型、可扩展、交互式的免费库，提供了一套查看卫星地图的 API
4、下面关于 Timetoast 的描述，哪个是错误的？(D)
A.Timetoast 是在线创作基于时间轴事件记载服务的网站
B.提供个性化的时间线服务

C.Timetoast 基于 flash 平台，可以在类似 flash 时间轴上任意加入事件
D.Timetoast 是一个提供复杂统计图表的工具
5、 下列关于可视化工具中高级分析工具的说法，错误的是？(B)
A.R 是属于 GNU 系统的一个自由、免费、源代码开放的软件
B.Weka 主要用于社交图谱数据可视化分析，可以生成非常酷炫的可视化图形
C.Gephi 主要用于社交图谱数据可视化分析，可以生成非常酷炫的可视化图形
D.R 通常用于大数据集的统计与分析

二、多选题

1、 在大数据时代，可视化技术可以支持实现哪些目标？（ABCD）
A.观测、跟踪数据
B.分析数据
C.辅助理解数据
D.增强数据吸引力
2、 信息图表是信息、数据、知识等的视觉化表达，下列哪个说法正确？(ABCD)
A.谷歌公司的制图服务接口 Google Chart API，可以用来为统计数据并自动生成图片
B.D3 是最流行的可视化库之一，是一个用于网页作图、生成互动图形的 JavaScript 函数库
C. ECharts 是由百度公司前端数据可视化团队研发的图表库，可以流畅地运行在 PC 和移动设备上
D.大数据魔镜是一款优秀的国产数据分析软件,它丰富的数据公式和算法可以让用户真正理解探索分析数据
3、 下列关于数据可视化的描述，正确的有？(ABC)
A.数据可视化是指将大型数据集中的数据以图形图像形式表示
B.数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示
C.利用数据分析和开发工具发现其中未知信息的处理过程
D.将数据的各个属性值以一维数据的形式表示
4、下列说法中，哪些是正确的？(ABCD)
A.Modest Maps 是一个小型、可扩展、交互式的免费库
B.Leaflet 是一个小型化的地图框架，通过小型化和轻量化来满足移动网页的需要
C.Google Fusion Tables 让一般使用者也可以轻松制作出专业的统计地图
D.大数据魔镜是一款优秀的国产数据分析软件,它丰富的数据公式和算法可以让用户真正理解探索分析数据

5、下面关于可视化图表的描述正确的是：（ABD）
A. 漏斗图适用于业务流程比较规范、周期长、环节多的流程分析
B. 树图是一种流行的、利用包含关系表达层次化数据的可视化方法
C. 桑基图是以特殊高亮的形式显示访客热衷的页面区域和访客所在的地理区域的图示
D. 词云对网络文本中出现频率较高的“关键词”给予视觉上的突出

第 9 章 大数据分析综合案例

本章无习题