大数据的主要特点包括五个V，即：挖掘性、速度、范围、质量和真实性。（）

【A.】√

【B.】×

【答案】B

Hadoop是一个开源的大数据处理工具，它的核心是MapReduce计算模型和HDFS分布式文件系统。（）

【A.】√

【B.】×

【答案】A

Apache Hadoop是一款支持MapReduce编程模型的开源框架，但它并不适用于处理大规模的数据集。（ ）

【A.】√

【B.】×

【答案】B

批处理是一种重要的大数据处理技术，它可以对静态数据进行批量处理，处理过程通常会涉及多个步骤，并将结果加载到分析数据存储中供分析工具使用和查询。（ ）

【A.】√

【B.】×

【答案】A

Hadoop生态系统中，Hive用于大规模数据集的抽取、转换和加载，而Pig用于数据分析。（）

【A.】√

【B.】×

【答案】B

Hadoop和Spark都是大数据处理的工具，其中Hadoop适用于处理大量静态数据，而Spark适合处理需要实时处理的数据流。（）

【A.】√

【B.】×

【答案】A

批处理系统支持作业执行状态的监控。（）

【A.】√

【B.】×

【答案】A

在MapReduce编程模型执行过程中，Reducing阶段由Reducer类负责，它将根据中间数据的键值对整理汇总后提供最终输出。（ ）

【A.】√

【B.】×

【答案】A

实时数据处理一般使用批处理模式来操作，通常会在数据生成后的几秒内完成处理。（）

【A.】√

【B.】×

【答案】B

MapReduce编程模型的首要步骤是对存储系统中的文件按列处理，并产生键值对。（）

【A.】√

【B.】×

【答案】B

实时处理系统和批处理系统相比，更能实现数据的实时性处理，但其对数据的处理质量和准确性有所损失。（）

【A.】√

【B.】×

【答案】A

大数据分析就是对大量复杂的数据进行分析，以揭示隐藏的模式、未知的关联、市场趋势、客户偏好等有用信息。（）

【A.】√

【B.】×

【答案】A

批量数据处理是一种在线计算模式，可以满足系统对数据进行实时处理的需要。（）

【A.】√

【B.】×

【答案】B

在MapReduce编程模型的执行流程中，数据洗牌/排序的工作是在输入阶段完成的。（ ）

【A.】√

【B.】×

【答案】B

MapReduce编程模型主要包括Map操作和Reduce操作两个阶段。（）

【A.】√

【B.】×

【答案】A

MapReduce可以通过在存储数据的节点上进行有效地任务调度来减少对网络资源的浪费，并且因为所有节点都在并行处理数据，可以大幅度缩短处理时间。（）

【A.】√

【B.】×

【答案】A

Hadoop是基于MapReduce编程模型开发的一个大数据处理框架。（）

【A.】√

【B.】×

【答案】A

批量数据处理旨在以最小延迟处理实时生成的数据流。（）

【A.】√

【B.】×

【答案】B

MapReduce基于分而治之的算法范式，利用多台计算机完成数据处理（）

【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。（）

【A.】√

【B.】×

【答案】A

大数据中的"大"主要指的是数据种类的多样性，而不是指数据量的大小。（ ）

【A.】√

【B.】×

【答案】B

大数据的重要特征之一是速度快。（ ）

【A.】√

【B.】×

【答案】A

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

DIKW模型描述了数据如何逐步转化为信息、知识及智慧的过程。（ ）

【A.】√

【B.】×

【答案】A

大数据的五个V特性是：体量（Volume）、速度（Velocity）、价值（Value）、多样性（Variety）和真实性（Veracity）。（）

【A.】√

【B.】×

【答案】A

结构化数据不可以用关系型数据库进行表示、存储与管理。（ ）

【A.】√

【B.】×

【答案】B

观测数据和试验数据的区别是前者在自然未被控制的条件下观测到，而后者是在有人工控制的条件下收集。（ ）

【A.】√

【B.】×

【答案】A

数据可视化的作用仅限于将抽象的数据进行可视表达，无法帮助我们发现数据中的规律和特征。（ ）

【A.】√

【B.】×

【答案】B

大数据处理的基本流程包括数据收集、数据预处理、数据存储与管理、数据分析与挖掘、数据展示五个环节。（ ）

【A.】√

【B.】×

【答案】A

大数据的一个重要特征是数据的多样性，这包括数据的各种格式和类型。（ ）

【A.】√

【B.】×

【答案】A

在大数据的处理过程中，数据清洗和数据整合是十分关键的步骤（）

【A.】√

【B.】×

【答案】A

大数据的生成主要来源于计算机系统的运行和人工输入，天气数据、网络数据、社交媒体数据等并不属于大数据的生成来源。（）

【A.】√

【B.】×

【答案】B

根据数据在收集过程中是否有控制因素，可以将数据分为截面数据和时间序列数据。（ ）

【A.】√

【B.】×

【答案】B

相对于结构化数据，非结构化数据的主要特点是没有固定的格式，它可能来自各种源，并且无法直接存入传统的关系型数据库。（ ）

【A.】√

【B.】×

【答案】A

在大数据的应用中，只需要依赖大数据分析技术，与各行业的业务关联性并不大。（）

【A.】√

【B.】×

【答案】B

在大数据的5V特性中，“可验证性”指的是数据的真实性和准确性，只有真实、准确的数据才能满足大数据处理和分析的需要。（ ）

【A.】√

【B.】×

【答案】B

办公文档、图像、声音和文本都是结构化数据。（ ）

【A.】√

【B.】×

【答案】B

根据数据在收集过程中是否控制有关因素，可以将数据分为观测数据和实验数据。（ ）

【A.】√

【B.】×

【答案】A

数据湖和数据仓库具有完全相同的特性和结构，可以互换使用。（ ）

【A.】√

【B.】×

【答案】B

大数据科学不仅包括数据采集和数据分析，还涵括数据预处理、数据存储和数据管理等环节。（）

【A.】√

【B.】×

【答案】A

研究数据仅仅包括数据采集和数据分析。（）

【A.】√

【B.】×

【答案】B

网页数据是一种半结构化数据。（）

【A.】√

【B.】×

【答案】A

大数据技术主要处理的数据类型是结构化数据，而对于非结构化数据处理能力较弱。（）

【A.】√

【B.】×

【答案】B

网页点击流数据是一种典型的非结构化大数据应用（）

【A.】√

【B.】×

【答案】A

截面数据是指同一时间不同观测个体的一个或多个指标组成的数据集。（ ）

【A.】√

【B.】×

【答案】A

时间序列数据是按时间顺序排列的观测值序列，用于所描述现象随时间变化的情况。（ ）

【A.】√

【B.】×

【答案】A

主成分分析方法无法用于降低数据的维度。（ ）

【A.】√

【B.】×

【答案】B

聚类分析与分类区别在于，分类需要预设定特定的类别，而聚类分析是无监督学习，无需预先设定类别，系统会自动将相似的数据聚集在一起，形成类簇。（）

【A.】√

【B.】×

【答案】A

异常检测是一种关键的大数据分析方法，可以用于发现数据中的异常或异常行为。（ ）

【A.】√

【B.】×

【答案】A

在进行数据分析时，数据预处理如数据清洗和数据转换是非必需的步骤。（ ）

【A.】√

【B.】×

【答案】B

决策树是一种简单高效的分类模型。（ ）

【A.】√

【B.】×

【答案】A

聚类分析仅能处理数值型数据，对于非数值型数据无效。（ ）

【A.】√

【B.】×

【答案】B

聚类分析是一种常用的无监督学习方法，用于发现数据集中的模式或结构。（ ）

【A.】√

【B.】×

【答案】A

在聚类的过程中，我们要事先知道数据中的类别信息，然后进行相应的分类。（ ）

【A.】√

【B.】×

【答案】B

决策树算法能够容忍数据中的缺失值和过滤存储在每个属性中的无关数据。（ ）

【A.】√

【B.】×

【答案】A

关联规则分析常用于购物篮分析，挖掘商品间的购买关系。（）

【A.】√

【B.】×

【答案】A

Excel是一种常用的电子表格程序，除了数据存储，还提供了丰富的数据分析和可视化功能。（ ）

【A.】√

【B.】×

【答案】A

对于预测样本的分类，决策树从根节点开始，根据预测样本的属性结果选择适当的分支，直到到达某个叶节点的类标号。（ ）

【A.】√

【B.】×

【答案】A

样本相关系数（correlation coefficient）可以用于衡量两个变量之间的线性关系强度。（ ）

【A.】√

【B.】×

【答案】A

k近邻法是基于示例的分类方法，根据最相似的样本的类别，来判断未知样本的类别。（）

【A.】√

【B.】×

【答案】A

k近邻算法实际上是一种分类算法，其通过找出样本最近的k个邻居并根据多数投票进行分类。（）

【A.】√

【B.】×

【答案】A

k近邻算法是一种简单但强大的聚类算法。（）

【A.】√

【B.】×

【答案】B

通过编写宏和VBA脚本，我们可以实现Excel自动化处理数据和复杂的数据分析任务。（ ）

【A.】√

【B.】×

【答案】A

在决策树算法中，每一次分枝的条件通常是将数据集的特征按某一指标进行划分，以此实现分类或预测任务。（）

【A.】√

【B.】×

【答案】A

统计学是以数据为研究对象的科学，以解决现实世界中的问题为目标。（ ）

【A.】√

【B.】×

【答案】A

在电子商务推荐算法中，基于关联规则的推荐算法是指根据商品本身属性计算商品之间的相似度。（）

【A.】√

【B.】×

【答案】B

时间序列分析中采用对数变换来消除异方差。（ ）

【A.】√

【B.】×

【答案】A

Excel只能存储数据，无法用于数据分析。（ ）

【A.】√

【B.】×

【答案】B

在HDFS中，NameNode为主节点，负责元数据管理和块位置的索引。（ ）

【A.】√

【B.】×

【答案】A

SQL数据库适合处理大量、高速和多样性的数据。（）

【A.】√

【B.】×

【答案】B

HDFS架构遵循主从结构，主节点称为数据节点，负责接收来自客户端的作业请求。（ ）

【A.】√

【B.】×

【答案】B

SQL数据库在定义关系时不需要定义属性的大小，从而避免了存储空间的浪费。（ ）

【A.】√

【B.】×

【答案】B

NoSQL数据库，如Cassandra和MongoDB，是设计用来处理大量数据的数据库系统。（）

【A.】√

【B.】×

【答案】A

关系型数据库不是用来存储和访问具有彼此相关性数据的数据库。（ ）

【A.】√

【B.】×

【答案】B

文件代表着程序与数据，它是位、字节、行或记录的序列。（ ）

【A.】√

【B.】×

【答案】A

在分布式文件系统HDFS中，数据丢失的问题可以通过在多个数据节点上复制相同的数据块来解决。（ ）

【A.】√

【B.】×

【答案】A

在大数据环境中，数据仓库主要用于存储结构化数据，而数据湖主要用于存储非结构化数据（ ）

【A.】√

【B.】×

【答案】B

当数据的结构发生变化时，使用SQL数据库进行相应调整的成本相对较低。（ ）

【A.】√

【B.】×

【答案】B

大数据处理需要将数据分布式存储在集群中的多台机器上进行并行处理，以正常运作。（ ）

【A.】√

【B.】×

【答案】A

在HDFS集群中，如果NameNode监测到某一个DataNode无法正常工作，它不会进行任务的重新分配。（ ）

【A.】√

【B.】×

【答案】B

大数据存储通常只需要考虑数据的存储，而不需要考虑数据的检索和处理。（）

【A.】√

【B.】×

【答案】B

关系数据库基于两个表中的公共列，可以建立两表之间的关系，避免数据的重复存储。（ ）

【A.】√

【B.】×

【答案】A

关系数据库是用来存储和访问具有彼此相关性数据的数据库。（ ）

【A.】√

【B.】×

【答案】A

NoSQL和NewSQL数据库是为满足非结构化数据和大数据处理应用的特殊要求，作为对关系型数据库的补充而出现的。（ ）

【A.】√

【B.】×

【答案】A

在HDFS架构中，主节点称为NameNode，而不是数据节点，它主要负责存储系统的元数据，处理客户端的请求。（ ）

【A.】√

【B.】×

【答案】A

NoSQL数据库是专门为大数据和实时网页应用设计的，因此通常不支持ACID事务。（ ）

【A.】√

【B.】×

【答案】A

NoSQL数据库如MongoDB和Cassandra不支持海量数据的存储，只能处理小规模数据。（ ）

【A.】√

【B.】×

【答案】B

HDFS系统通过在不同节点上存储相同数据块的备份，来实现对文件的高效并行处理和数据的容错性。（ ）

【A.】√

【B.】×

【答案】A

面部识别技术的快速发展并没有带来任何隐私侵权的问题。（ ）

【A.】√

【B.】×

【答案】B

黑客技术的非法入侵是大数据隐私保护所面临的主要威胁之一。（）

【A.】√

【B.】×

【答案】A

在大数据环境下，数据使用的合规性并不需要考虑，因为大数据的收集和分析主要是为了提供更好的服务和产品。（ ）

【A.】√

【B.】×

【答案】B

对于存储的敏感信息，完全脱敏后的信息仍然可以进行有效的数据分析和挖掘。（ ）

【A.】√

【B.】×

【答案】A

为防止隐私被窃取，企业和个人可以运用隐私保护技术来保护数据的安全。（ ）

【A.】√

【B.】×

【答案】A

适当地使用数据脱敏技术，可以有效地减少敏感数据在采集、传输、使用等环节中的暴露，降低敏感数据泄露的风险。（）

【A.】√

【B.】×

【答案】A

数据安全性与隐私保护并不需要遵守相关法律和政策规定。（）

【A.】√

【B.】×

【答案】B

匿名化处理是一种处理个人识别信息以防止个人被识别的技术，它是大数据隐私保护的重要手段。（ ）

【A.】√

【B.】×

【答案】A

数据脱敏技术无法降低敏感数据泄露的风险。（）

【A.】√

【B.】×

【答案】B

遵守GDPR等数据保护法规，不仅能够保护个人和企业数据的安全，也有助于提升企业的声誉和客户信任。（）

【A.】√

【B.】×

【答案】A

数据脱敏技术，数据加密技术等都能有效保护大数据的安全。（ ）

【A.】√

【B.】×

【答案】A

大数据技术中数据的加密是一种有效的保护数据隐私的方法，它可以有效防止数据在传输过程中被截获。（）

【A.】√

【B.】×

【答案】A

在大数据环境中，数据加密是主要的安全处理技术，因为这可以确保即使数据泄露，也无法被滥用。（）

【A.】√

【B.】×

【答案】A

随着大数据平台的广泛使用，对隐私信息的保护变得无关紧要。（）

【A.】√

【B.】×

【答案】B

数据脱敏技术可以有效降低敏感数据泄露的风险。（）

【A.】√

【B.】×

【答案】A

大数据在教育领域的应用主要包括个性化教学和学生行为分析等。（）

【A.】√

【B.】×

【答案】A

大数据在诸如消费者行为监测、市场趋势分析等一系列商业领域都有广泛的应用。（ ）

【A.】√

【B.】×

【答案】A

疾病预防和治疗属于大数据在生物医学领域的应用。（ ）

【A.】√

【B.】×

【答案】A

大数据无法帮助理解疾病的诱因和预防措施，只能在疾病发生后进行诊断。（ ）

【A.】√

【B.】×

【答案】B

大数据对于环保领域没有实际应用价值。（ ）

【A.】√

【B.】×

【答案】B

通过收集特定疾病患者群体的数据信息来分析该疾病的具体诱因，是大数据在医学领域中的一种重要应用。（ ）

【A.】√

【B.】×

【答案】A

大数据技术仅限于科技领域的应用，对于传统领域如农业、教育等没有实际应用价值。（ ）

【A.】√

【B.】×

【答案】B

基于用户的协同过滤推荐算法是根据商品或内容的点赞、收藏、分享等情况判断用户对该商品的兴趣和偏好程度，并推荐相似商品给用户。（ ）

【A.】√

【B.】×

【答案】B

银行业通过大数据技术可以有效分析经营过程中可能存在的风险因素。（）

【A.】√

【B.】×

【答案】A

银行业利用大数据技术可以分析客户全生命周期的不同需求，从而为客户提供更满意、更精准的产品与服务。（ ）

【A.】√

【B.】×

【答案】A

大数据技术在医疗领域无法帮助医生进行疾病诊断和预防。（）

【A.】√

【B.】×

【答案】B

电子商务通过大数据技术优化推荐算法，改善用户购物体验。（ ）

【A.】√

【B.】×

【答案】A

在电商行业，通过大数据分析用户行为，可以提升产品销售、优化营销策略并改善用户体验。（）

【A.】√

【B.】×

【答案】A

将深度学习引入医学影像大数据平台有助于提高医学影像筛查诊断的准确率和自动化程度。（ ）

【A.】√

【B.】×

【答案】A

只能通过数据插补的方式处理数据集中的缺失值。（ ）

【A.】√

【B.】×

【答案】B

大数据预处理阶段不包括数据插补步骤。（ ）

【A.】√

【B.】×

【答案】B

在大数据分析过程中，无论数据采集还是预处理都是不可或缺的环节，它们决定了分析结果的质量。（ ）

【A.】√

【B.】×

【答案】A

数据采集技术的质量并不影响数据的采集效率和数据质量。（ ）

【A.】√

【B.】×

【答案】B

数据融合是数据预处理的一种方式，它主要用于整合来自多个数据源的信息，从而在同一视角下进行数据处理和分析。（ ）

【A.】√

【B.】×

【答案】A

缺失值的处理方法只有直接删除，没有其他的处理方法。（ ）

【A.】√

【B.】×

【答案】B

大数据的采集技术已经逐渐成为数据高速增长的主要推动因素之一，成为大数据技术的重要活动。（ ）

【A.】√

【B.】×

【答案】A

数据预处理的一种方法是数据归一化，它将所有数据缩放到一个固定的范围中，以消除数据之间的尺度差异。（ ）

【A.】√

【B.】×

【答案】A

数据清洗包括去除重复数据、填补缺失值以及修正明显的错误值等步骤。（ ）

【A.】√

【B.】×

【答案】A

特征选择是指选择对预测结果影响力较大的特征，以改善学习算法的性能和解释性。（ ）

【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。（）

【A.】√

【B.】×

【答案】A

在大数据的应用中，数据的采是非必须的步骤。（ ）

【A.】√

【B.】×

【答案】B

回归插补通过建立缺失变量和已观测变量之间的回归方程，用缺失变量的预测值进行缺失值的插补。（ ）

【A.】√

【B.】×

【答案】A

数据预处理的主要目的是为了提高数据质量，将原始数据变成更加方便计算或处理的格式，使数据形态更加符合建模要求，进而提升数据挖掘的质量和效率。（ ）

【A.】√

【B.】×

【答案】A

数据缺失的现象大量存在而又无法避免。（ ）

【A.】√

【B.】×

【答案】A

数据预处理的过程中不需要进行数据转换。（ ）

【A.】√

【B.】×

【答案】B

数据预处理过程中的数据清洗包括去除重复的数据，那些离群点无需进行处理。（ ）

【A.】√

【B.】×

【答案】B

对于数据缺失的现象，数据预处理通过如删除、插补等方式进行处理。（ ）

【A.】√

【B.】×

【答案】A

数据清洗就是将无用的、重复的、不完整的数据进行修改或删除，以保证数据质量的过程。（ ）

【A.】√

【B.】×

【答案】A

为了避免数据的杂乱无序，通常将数据整理成数据集。（ ）

【A.】√

【B.】×

【答案】A

数据采集，也叫“数据获取”，就是从真实世界中采集和记录信息，将信息数据化形成原始数据。（ ）

【A.】√

【B.】×

【答案】A

数据的收集过程中，数据采集的工具和方法对数据质量、全面性和后续处理没有影响。（ ）

【A.】√

【B.】×

【答案】B

数据降维是一种能有效减少数据复杂度，简化模型计算的技术。（ ）

【A.】√

【B.】×

【答案】A

特征选择是指由原始变量的函数构造一些新的变量，新的变量能保留原始变量的绝大部分信息。（）

【A.】√

【B.】×

【答案】B

数据可视化可以帮助用户更快速地理解和掌握数据的含义、结构和重要特性。（ ）

【A.】√

【B.】×

【答案】A

对于一个分类变量的各类别的频数占比，常用柱形图、条形图和饼图来展示。（）

【A.】√

【B.】×

【答案】A

在数据分析中，数据可视化只是一种辅助工具，不具有实际的分析功能。（ ）

【A.】√

【B.】×

【答案】B

散点图是最常用于展示两个数值型变量之间关系的可视化方法。（ ）

【A.】√

【B.】×

【答案】A

雷达图适用于反映多个变量的相对大小与差异，不适合反映变量之间的相互关系。（ ）

【A.】√

【B.】×

【答案】A

直方图是用来显示一个变量的分布情况，不适用于展现两个变量之间的关系。（ ）

【A.】√

【B.】×

【答案】A

气泡图中气泡的面积大小没有实际意义。（ ）

【A.】√

【B.】×

【答案】B

直方图是将数据点用直线串联起来所形成的统计图（）

【A.】√

【B.】×

【答案】B

对于一组数值型数据的中位数、四分位数等分布情况，常用盒形图来展示。（）

【A.】√

【B.】×

【答案】A

通常使用直方图展示两个数值型变量之间的相关关系。（ ）

【A.】√

【B.】×

【答案】B

折线图适合用于展示时间序列数据的长期趋势。（）

【A.】√

【B.】×

【答案】A

数据可视化对于提升数据的理解、分析与推断没有帮助。（ ）

【A.】√

【B.】×

【答案】B

在数据可视化技术中，我们使用计算机图形学、图像处理和人工交互等技术将数据转化为图像等可视形式。 （ ）

【A.】√

【B.】×

【答案】A

在Hadoop生态系统中，以下哪个组件是分布式文件系统，主要用于存储大数据（ ）

【A.】HDFS

【B.】MapReduce

【C.】YARN

【D.】Storm

【答案】A

在Hadoop生态系统中，主要负责节点集群的任务调度和资源分配，将存储和计算资源分配给不同应用程序的组件是（ ）

【A.】HDFS

【B.】MapReduce

【C.】YARN

【D.】Storm

【答案】C

Apache Spark是一个（ ）

【A.】数据库系统

【B.】编程语言

【C.】大数据处理框架

【D.】机器学习库

【答案】C

HDFS的高可用性是指（ ）

【A.】随着需求的增加，集群可以轻松扩展到更多节点

【B.】实现节点集群上的并行数据处理

【C.】即使集群中的某个节点发生故障，数据仍然可用

【D.】出现故障时，可以从集群中的其他节点获取数据备份

【答案】C

下列各项不属于云计算特点的是（ ）

【A.】敏捷

【B.】弹性计算

【C.】节约成本

【D.】过度配置资源

【答案】D

HDFS的可扩展性主要体现在（）

【A.】随着需求的增加，集群可以轻松扩展到更多节点

【B.】即使集群中的某个节点发生故障，数据依然可用

【C.】以分布式方式存储数据，减少处理时间

【D.】确保数据始终可用，防止数据丢失

【答案】A

下列哪一项不是实时数据处理系统设计的核心问题？

【A.】实现信息的实时获取、处理和存储

【B.】处理过程不能影响信息的获取

【C.】数据存储系统需要支持大量的写入操作

【D.】必须定时备份数据

【答案】D

下列关于Hadoop框架的描述中，错误的是（）

【A.】Hadoop是一个开源的分布式计算框架

【B.】Hadoop的核心是HDFS和MapReduce

【C.】Hadoop无法处理大数据

【D.】Hadoop可以在廉价的硬件上运行

【答案】C

在Hadoop生态系统中，主要负责支持数据查询和数据分析，并通过SQL-like语言HiveQL实现结构化查询的组件是（ ）

【A.】Hive

【B.】HiveQL

【C.】HDFS

【D.】Storm

【答案】A

关于Storm框架的描述，下列哪一项是不正确的？

【A.】Storm是具有高度容错性、灵活性、可靠性的处理系统

【B.】Storm允许实时流处理

【C.】Storm的延迟较高

【D.】Storm具有高度可扩展性。

【答案】C

下列各项不属于批处理系统的特点的是（ ）

【A.】为开发者提供了一个简单、快捷的开发框架

【B.】支持各种数据格式的处理

【C.】支持数据在不同系统之间进行交换

【D.】可以实现实时的分析报告或自动响应

【答案】D

在工业网络实时监控系统中，需要连续不断地采集和处理数据。以下（ ）不属于这种计算模式

【A.】在线处理

【B.】实时处理

【C.】流式计算

【D.】批量计算

【答案】D

在大数据处理过程中，Map-Reduce的主要应是（ ）

【A.】交互式查询

【B.】流式处理

【C.】批处理

【D.】实时处理

【答案】C

在大数据处理技术中，下列哪一项是用于进行大规模数据流处理的框架（ ）

【A.】Hadoop

【B.】MapReduce

【C.】YARN

【D.】Storm

【答案】D

在大数据处理中，下列哪种系统可以支持实时分析报告或自动响应（）

【A.】批处理系统

【B.】实时处理系统

【C.】离线处理系统

【D.】Hadoop系统

【答案】B

在Hadoop生态系统中，主要负责提供实时处理和分析流数据的组件是（）

【A.】HDFS

【B.】MapReduce

【C.】YARN

【D.】Storm

【答案】D

下列各项不属于Hadoop的特点是（ ）

【A.】存储迅速

【B.】成本高

【C.】计算能力强

【D.】灵活性强

【答案】B

下列哪项不是Spark的主要特点（ ）

【A.】快速数据处理

【B.】内存计算

【C.】无法处理复杂数据

【D.】容错性强

【答案】C

下列哪一项大数据处理技术是流数据处理框架（）？

【A.】Spark

【B.】Storm

【C.】HDFS

【D.】MapReduce

【答案】B

在Hadoop生态系统中，被视作Hadoop生态系统的大脑，负责节点集群的任务调度和资源分配的组件是（ ）

【A.】HBase

【B.】HDFS

【C.】YARN

【D.】Pig

【答案】C

下列哪一项不是Spark与Hadoop相比的优势（）？

【A.】同时支持批处理和流处理

【B.】算法复杂

【C.】内存存储，处理速度快

【D.】支持多种语言接口

【答案】B

下列不属于Storm框架特点的是（）

【A.】允许实时流处理

【B.】具有高度可扩展性

【C.】延迟高

【D.】支持多种编程语言

【答案】C

Hadoop生态系统是一种（ ）

【A.】编程语言

【B.】服务

【C.】统计分析软件

【D.】解决大数据问题的平台

【答案】D

Hadoop生态系统中的（ ）负责节点集群的任务调度和资源分配

【A.】HDFS

【B.】YARN

【C.】MapReduce

【D.】Hive

【答案】B

在Hadoop生态系统中，主要负责跨节点存储结构化或非结构化数据，并以日志文件的形式管理数据的组件是（ ）

【A.】HDFS

【B.】MapReduce

【C.】YARN

【D.】Storm

【答案】A

下列各项不属于批处理系统的特点的是（ ）

【A.】可以实现实时的分析报告或自动响应

【B.】可以实现无缝扩展以处理峰值数据量或数据请求

【C.】支持数据在不同系统之间进行交换

【D.】支持作业执行状态的监控

【答案】A

在MapReduce模型编程中，运算步骤不包括（）

【A.】遍历数据

【B.】映射键值

【C.】数据分组

【D.】数据插补

【答案】D

在以二维表形式表示的数据集中，列表示（ ）

【A.】特征

【B.】样本

【C.】属性

【D.】字段

【答案】A

在大数据5V特征中，“Veracity”表示的是（ ）

【A.】体量大

【B.】种类多

【C.】真实性

【D.】价值大

【答案】C

下列哪种数据属于半结构化数据（）

【A.】XML

【B.】关系数据库中的数据

【C.】图像

【D.】二维数据表

【答案】A

下列哪项不是大数据存储技术（）

【A.】NoSQL数据库

【B.】Hadoop

【C.】MapReduce

【D.】Excel

【答案】D

大数据处理的核心环节是（ ）

【A.】数据的采

【B.】数据的存储与管理

【C.】数据的可视化

【D.】数据的分析与挖掘

【答案】D

下列各项属于定量变量的是（ ）

【A.】性别

【B.】婚姻状况

【C.】年龄

【D.】国籍

【答案】C

下列关于数据类型描述正确的是（）

【A.】观测数据是在有人工控制的条件进行观察和收集的数据

【B.】时间序列数据和截面数据都是按时间顺序排列的观测值序列

【C.】非结构化数据是可以使用关系型数据库进行表示、存储与管理的数据

【D.】半结构化数据是处于结构化数据和非结构化数据之间的数据

【答案】D

在大数据分析中，以下哪种分析是用来挖掘数据中的隐藏模式和关联的？

【A.】描述性分析

【B.】诊断性分析

【C.】预测性分析

【D.】探索性分析

【答案】D

下列关于数据类型的描述中，哪一项是错误的（）？

【A.】非结构化数据包括文本、图像和音视频等

【B.】结构化数据指的是预定格式的数据如数据库

【C.】半结构化数据包括XML文件和电子邮件

【D.】非结构化数据易于进行大数据分析

【答案】D

下面哪项不属于大数据预处理步骤（）

【A.】数据清洗

【B.】数据集成

【C.】数据转换

【D.】数据安全

【答案】D

以下哪项不属于大数据处理的步骤？

【A.】数据的获取

【B.】数据的分析

【C.】数据的保存

【D.】数据的丢弃

【答案】D

在数据集中，行常常表示（ ）

【A.】属性

【B.】样本

【C.】文件

【D.】数据集

【答案】B

在大数据的处理过程中，以下哪项不是数据准备阶段的任务（）？

【A.】数据集成

【B.】数据清洗

【C.】数据挖掘

【D.】数据变换

【答案】C

下列关于非结构化数据的描述，哪一项是错误的（）

【A.】非结构化数据没有固定的数据模型

【B.】非结构化数据包括图像、声音、视频等

【C.】非结构化数据不可以进行数据分析

【D.】非结构化数据数量大

【答案】C

下列哪项技术不是用于处理大数据的（）？

【A.】实时计算技术

【B.】分布式存储技术

【C.】并行计算技术

【D.】高温超导技术

【答案】D

大数据5V特征中的Veracity表示（ ）

【A.】体量大

【B.】种类多

【C.】真实性

【D.】数据快

【答案】C

以下哪一项不是结构化数据（）

【A.】关系数据库中的数据

【B.】Excel表格中的数据

【C.】歌词文本

【D.】二维数据表

【答案】C

大数据分析的主要目标不包括（）

【A.】预测未来走势

【B.】抽取信息

【C.】描绘数据形状

【D.】隐藏数据

【答案】D

下列哪项不是二手数据的特点?（）

【A.】数据不准确

【B.】采集成本高

【C.】与研究目的相关度不够

【D.】口径可能不一致

【答案】B

关于时间序列数据的描述，以下哪一项是错误的？

【A.】是按时间顺序排列的观测值序列

【B.】不能使用统计方法进行分析

【C.】在时间间隔上接近的观测值通常更相关

【D.】两个观测值在时间间隔上越远，它们的相关性就更弱

【答案】B

关于传统数据与大数据的获取与处理，下列描述中错误的是（）

【A.】传统的数据获取来源单一

【B.】采用关系型数据库和并行数据库可以处理传统数据

【C.】获取大数据的数据来源广泛

【D.】大数据的数据量相对较小

【答案】D

下列哪一项不是定性变量的例子（）？

【A.】职业

【B.】性别

【C.】国籍

【D.】身高

【答案】D

下列哪一项不是大数据处理面临的挑战（）

【A.】数据安全和隐私保护

【B.】高效实时处理大规模数据

【C.】从大量数据中发现有用信息和知识

【D.】数据量小导致结果不准确

【答案】D

根据原始数据是否为数据的直接来源，可以将数据分为（ ）

【A.】结构化、非结构化和半结构化数据

【B.】一手数据和二手数据

【C.】观测数据和实验数据

【D.】截面数据和时间序列数据

【答案】B

下列哪一项是定性变量（）

【A.】年龄

【B.】股票价格

【C.】数学成绩

【D.】信用等级

【答案】D

关于大数据，以下哪个描述是正确的？

【A.】大数据只包括结构化数据

【B.】大数据只关乎数据的体积

【C.】大数据处理没有安全隐私问题

【D.】大数据可以来自各种来源，包括业务交易、社交媒体等

【答案】D

当一个变量只能取有限几个值时，该变量就是（ ）

【A.】定量变量

【B.】定性变量

【C.】独立变量

【D.】连续变量

【答案】B

数据预处理的主要目的是什么?

【A.】让数据产生价值

【B.】提高数据处理的效率

【C.】解决数据缺失、不一致、冗余等问题

【D.】将数据转化为图形

【答案】C

按照数据的结构化程度，我们可将数据分为（ ）

【A.】截面数据和时间序列数据

【B.】一手数据和二手数据

【C.】观测数据和实验数据

【D.】结构化数据、半结构化数据和非结构化数据

【答案】D

下列关于大数据的描述中，哪一项是错误的（）

【A.】大数据来源广泛

【B.】大数据处理需要复杂的计算模型

【C.】大数据总体分析结果总是准确的

【D.】大数据中包含大量废弃的、质量低下的数据

【答案】C

在以二维表形式表示的数据集中，行表示（ ）

【A.】特征

【B.】样本

【C.】属性

【D.】字段

【答案】B

在进行数据分析时，通常将定性变量转换为（）以便进行数学分析和建模

【A.】定量变量

【B.】分类变量

【C.】矩阵

【D.】函数

【答案】A

下列哪一个不是定量变量（）

【A.】体重

【B.】身高

【C.】教育程度

【D.】薪资

【答案】C

下列各项属于非结构化数据的是（ ）

【A.】图像

【B.】二维数据表

【C.】HTML文档

【D.】以上均是

【答案】A

关于定量变量和定性变量的描述中，哪一项是错误的（）

【A.】定量变量具有度量的含义

【B.】定性变量是不能用数值带量的

【C.】性别是定量变量

【D.】年龄是定量变量

【答案】C

下列哪种方法不属于大数据采集方法（）？

【A.】无线传感网络

【B.】日志文件

【C.】互联网爬虫

【D.】数据建模

【答案】D

下列属于图数据的主要特性的是（）

【A.】数据驱动计算

【B.】不规则问题

【C.】高数据访问率

【D.】以上均是

【答案】D

下列各项不属于大数据特征的是（ ）

【A.】体量大

【B.】种类多

【C.】真实性

【D.】数据生成慢

【答案】D

下列各项属于定性变量的是（ ）

【A.】收入

【B.】支出

【C.】身高

【D.】印象

【答案】D

数据集是一个（ ）

【A.】文本

【B.】文件

【C.】图像

【D.】图片

【答案】B

下面不属于大数据的处理过程的是（）

【A.】数据获取

【B.】数据清洗

【C.】数据分析

【D.】数据安全

【答案】D

当一个变量可以在一个范围内连续取值时，该变量就是（ ）

【A.】定量变量

【B.】定性变量

【C.】属性变量

【D.】分类变量

【答案】A

在大数据的处理流程中，下列各项中最先进行的是（）

【A.】存储与管理

【B.】可视化

【C.】采

【D.】分析与挖掘

【答案】C

下列各项属于结构化数据的是（ ）

【A.】图像

【B.】二维数据表

【C.】声音

【D.】文本

【答案】B

下列各项属于大数据的成因因素的是（ ）

【A.】数据储存能力

【B.】数据管理能力

【C.】数据采集能力

【D.】以上皆是

【答案】D

传统的主要基于统计学的分析方法在大数据时代所面临的挑战包括（ ）

【A.】全体数据，不是随机样本

【B.】放弃数据的精确性，尽可能收集更多数据

【C.】重视相关关系，而不是因果关系

【D.】以上均是

【答案】D

在k近邻算法中，对于未知样本的类别判定，我们依据的原则是（ ）

【A.】最短距离

【B.】最长距离

【C.】距离平均值

【D.】近邻多数类别判定

【答案】D

以下哪种大数据分析方法旨在将相似对象组合在一起？

【A.】聚类分析

【B.】关联规则分析

【C.】决策树分析

【D.】线性回归分析

【答案】A

下面哪不适合用k近邻法进行分类（ ）

【A.】良恶性肿瘤预测

【B.】天气预报

【C.】人脸识别

【D.】邮件分类

【答案】B

关于相关系数的信息，以下哪项说法是错误的？

【A.】相关系数用于衡量两个变量间的线性关系强度

【B.】根据总体全部数据计算的相关系数称为总体相关系数

【C.】相关系数的值介于0到1之间

【D.】相关系数也被称为线性相关系数和皮尔森相关系数

【答案】C

下列哪项是无监督学习的典型应用？（ ）

【A.】聚类分析

【B.】回归

【C.】分类

【D.】以上均是

【答案】A

下列哪种数据挖掘方法不可用于异常值检测（ ）

【A.】Z-score方法

【B.】基于距离的方法

【C.】基于密度的方法

【D.】逻辑回归

【答案】D

聚类分析在大数据分析中的主要功能是什么？（）

【A.】通过数据自动发现潜在的类

【B.】分析数据的分布特征

【C.】预测未来的数据变化

【D.】测量数据的准确性和稳定性

【答案】A

根据已知的类标号建立分类模型的数据集是（ ）

【A.】训练集

【B.】检验集

【C.】预测集

【D.】测试集

【答案】A

以下关于聚类分析的描述中，错误的是（ ）

【A.】聚类分析是一种无监督学习方法

【B.】K-means是一种常用的聚类分析方法

【C.】聚类分析的目的是找出已知类别的数据

【D.】聚类分析常用于市场分析和社交网络分析

【答案】C

下列关于推断统计学的描述中，错误的是（ ）

【A.】参数估计是根据样本的统计量来估计总体中的参数

【B.】假设检验可以度量变量之间的相关程度

【C.】判别分析是将某个对象归到已知类别中

【D.】时间序列分析是研究时间序列数据变化规律的方法

【答案】B

在大数据的挖掘中，用于解决分类问题的机器学习算法不包括（ ）

【A.】决策树

【B.】支持向量机

【C.】k-近邻

【D.】线性回归

【答案】D

假设散点图中的观测点分布较为分散，没有任何规律，说明两个变量之间的关系为（）

【A.】完全线性相关

【B.】线性相关

【C.】非线性相关

【D.】不相关

【答案】D

下列各项关于分类的描述中，错误的是（）

【A.】可以借助分类方法根据电子邮件的标题判断其是否为垃圾邮件

【B.】在进行建模之前就要有明确的分组预测目标

【C.】k近邻算法是一种简单但强大的分类算法

【D.】用来建立分类模型的输入数据称为测试集

【答案】D

下列各项不属于推断统计学的是（ ）

【A.】参数估计

【B.】描述统计分析

【C.】假设检验

【D.】方差分析

【答案】B

以下哪种方法可以用于对数据集合进行降维处理（ ）

【A.】主成分分析

【B.】K-means聚类

【C.】支持向量机

【D.】关联规则

【答案】A

下列各项属于机器学习算法的是（ ）

【A.】监督学习

【B.】无监督学习

【C.】半监督学习

【D.】以上均是

【答案】D

在大数据分析中，下列哪种语言拥有丰富的数据分析库和机器学习库？

【A.】Java

【B.】Python

【C.】C++

【D.】JavaScript

【答案】B

下列哪种方法不能用于解决分类问题（ ）

【A.】决策树

【B.】感知器

【C.】K-means聚类

【D.】随机森林

【答案】C

下列哪项不是聚类分析的功能或用途？（）

【A.】识别出那些不同于已知类别的异常行为

【B.】确定保单持有者的分类

【C.】对搜索结果进行自动聚类分析

【D.】协助完成预测分析

【答案】D

以下关于基于关联规则的推荐算法的描述中，正确的是（ ）

【A.】根据用户对商品或内容的收藏或分享等情况判断用户对该商品的兴趣和偏好程度

【B.】根据用户的基本信息发现用户之间的相似情况，将相似用户喜爱的的其他商品推荐给当前用户

【C.】挖掘不同商品在销售过程中的相关性

【D.】根据商品本身的属性数据，计算商品之间的相似度，再基于用户的历史浏览等信息推荐给用户相似的商品

【答案】C

在分类效果的评估中，何为"过拟合"现象？

【A.】训练集和测试集的表现相同

【B.】训练集和测试集的表现完全不同

【C.】训练集性能远超测试集

【D.】测试集性能远超训练集

【答案】C

在数据挖掘中，用以预测具有连续值输出的问题，一般选用（ ）方法

【A.】分类

【B.】聚类

【C.】回归

【D.】关联规则挖掘

【答案】C

下列哪个不是机器学习的主要任务（ ）

【A.】分类

【B.】回归

【C.】聚类

【D.】数据存储

【答案】D

下列哪种语言被广泛应用于大数据分析和机器学习中（ ）

【A.】Java

【B.】Python

【C.】C++

【D.】Javascript

【答案】B

以下哪种方法不属于常见的数据挖掘任务（）

【A.】分类

【B.】聚类

【C.】关联规则挖掘

【D.】数据标记

【答案】D

假设散点图中的观测点恰好落在一条直线上，说明两个变量之间的关系为（）

【A.】完全线性相关

【B.】线性相关

【C.】非线性相关

【D.】不相关

【答案】A

k近邻算法是一种简单但强大的（ ）算法

【A.】分类

【B.】聚类

【C.】相关分析

【D.】回归分析

【答案】A

下列各项属于数据分类算法的是（ ）

【A.】k近邻算法

【B.】决策树

【C.】支持向量机

【D.】以上均是

【答案】D

下列各项关于聚类的描述中，错误的是（）

【A.】可以借助聚类方法进行异常检测

【B.】在进行建模之前就要有明确的分组预测目标

【C.】可以利用聚类分析发现具有相似功能的基因组

【D.】根据数据本身的自然结构对数据进行分组

【答案】B

以下关于关联规则的描述中，错误的是（ ）

【A.】可以用于预测商品的价格

【B.】可以用于找到商品之间的关联

【C.】可以用于购物篮数据分析

【D.】可以用于电商的推荐系统

【答案】A

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

在大数据分析和处理中，以下哪种算法经常用于进行数据降维？

【A.】决策树算法

【B.】贝叶斯算法

【C.】随机森林算法

【D.】主成分分析算法

【答案】D

以下哪种大数据分析技术可以发现数据集中潜在的关联规则？

【A.】聚类分析

【B.】决策树

【C.】关联规则挖掘

【D.】线性回归

【答案】C

世界上最早采用图形菜单驱动界面的统计分析软件是（ ）

【A.】Excel

【B.】SPSS

【C.】R语言

【D.】Python语言

【答案】B

根据概率的原则进行分类的机器学习算法是（ ）

【A.】k近邻算法

【B.】决策树

【C.】朴素贝叶斯分类器

【D.】随机森林

【答案】C

下列各项不能用于描述数据集中趋势的是（ ）

【A.】方差

【B.】平均数

【C.】中位数

【D.】峰值

【答案】D

在分类效果的评估中，使用（ ）测度指标来表示不正确分类的比例

【A.】正确率

【B.】错误率

【C.】召回率

【D.】特异性

【答案】B

下列关于数据湖的描述中，错误的是（ ）

【A.】数据湖通常存储不经过清洗和整理的原始数据

【B.】数据湖支持自助式的数据查询和分析

【C.】数据湖的数据通常经过严格的模式定义和结构化清洗

【D.】数据湖能够处理多种来源，不同类型的数据

【答案】C

与管理

在进行海量数据的存储时，较为适合的数据库类型是（ ）

【A.】关系型数据库

【B.】面向对象数据库

【C.】分布式数据库

【D.】层次数据库

【答案】C

与管理

以下哪项不是大数据存储技术Hadoop的主要组成部分（）？

【A.】HDFS

【B.】MapReduce

【C.】HBase

【D.】Photoshop

【答案】D

与管理

关于NoSQL数据库的描述中，以下哪项是错误的（）？

【A.】适合存储大量数据

【B.】无需预定义数据模型

【C.】支持ACID事务

【D.】具有高扩展性

【答案】C

与管理

数据仓库与传统数据库的最大区别是什么？

【A.】数据仓库不包含数据库

【B.】数据仓库主要用于数据分析而非数据存储

【C.】数据仓库只存储历史数据

【D.】数据仓库只能在云平台上部署

【答案】B

与管理

下列有关SQL数据库的局限性，描述错误的是（）

【A.】结构化数据建模

【B.】已有数据结构的变化成本高

【C.】空间利用率低

【D.】无法保证事务处理的可靠性

【答案】D

与管理

下列关于文档数据库的描述中，正确的是（ ）

【A.】将数据存储在表、行和动态列

【B.】每个文档都包含成对的字段和值

【C.】将数据存储在节点和边

【D.】使用哈希表，表中包含键和用来指向特定数据的指针

【答案】B

与管理

在HDFS的工作机制中，如果NameNode无法正常工作，它会做什么？（）

【A.】停止所有在DataNode上的作业

【B.】重新分配数据块

【C.】重新分配DataNode的处理任务

【D.】将数据块移动到另一个节点

【答案】C

与管理

目前处理大数据的需求，更倾向于使用哪种数据库？（）

【A.】SQL数据库

【B.】NoSQL数据库

【C.】Microsoft Access数据库

【D.】Oracle数据库

【答案】B

与管理

下列哪项不是大数据存储技术NoSQL的特点（）？

【A.】高并发读写

【B.】建立严格的数据关系结构

【C.】易扩展

【D.】处理海量数据的能力

【答案】B

与管理

下列关于数据仓库的特点，描述错误的是（ ）

【A.】数据以主题为导向，提供决策支持

【B.】关注数据随时间的变化，支持历史数据分析

【C.】数据质量高，提供数据一致性和准确性

【D.】数据源和数据类型单一

【答案】D

与管理

下列关于NoSQL数据库的描述中，哪项是错误的（）？

【A.】适合处理大规模数据

【B.】支持SQL查询语言

【C.】不需要预定义结构

【D.】各种类型的NoSQL数据库适用于不同的需求

【答案】B

与管理

在Hadoop的HDFS架构中，以下哪项功能不是主节点（NameNode）的职责（）？

【A.】管理文件系统命名空间

【B.】维护文件系统树及元数据

【C.】存储实际的数据块

【D.】管理文件系统的客户端应用

【答案】C

与管理

在大数据存储与管理中，下列哪项技术不属于Hadoop生态圈？

【A.】Hive

【B.】Pig

【C.】HBase

【D.】Redis

【答案】D

与管理

在数据仓库的三层架构中，哪一层可以实现数据的快速查询？

【A.】顶层

【B.】中间层

【C.】底层

【D.】不能实现快速查询

【答案】B

与管理

HDFS中，如何确保数据的可靠性？

【A.】不断增加数据的存储空间

【B.】对数据进行冗余存储，即保存多个数据备份，并将这些备份存储到不同的节点

【C.】将所有数据存储在一台计算机服务器上

【D.】频繁清理数据

【答案】B

与管理

下列各项不属于NoSQL数据库的是（ ）

【A.】关系数据库

【B.】文档数据库

【C.】键值存储数据库

【D.】图形数据库

【答案】A

与管理

数据仓库为什么可以帮助企业保持竞争力？

【A.】提供数据分析功能以及数据可视化和演示功能

【B.】数据仓库可以自动清理垃圾数据

【C.】使用数据仓库可以节省存储空间

【D.】数据仓库提供了大量质量低下的数据供企业使用

【答案】A

与管理

下列哪一项不是数据湖的特点（ ）

【A.】存储所有类型的数据，包括结构化、半结构化和非结构化数据

【B.】数据以预定义的模式存储

【C.】原始数据保留，不会过早转化

【D.】支持大规模的数据科学和机器学习工作负载

【答案】B

与管理

下列各项属于数据仓库的特点的是（ ）

【A.】数据质量不高

【B.】关注数据随时间的变化

【C.】只接受同类型数据

【D.】无法提供决策支持

【答案】B

与管理

下列各项不属于数据仓库的特点的是（）

【A.】将分析处理与事务数据库分离

【B.】整合来自不同数据源、不同类型的数据

【C.】数据质量高

【D.】数据仓库是面向应用而设计的

【答案】D

与管理

在HDFS架构中，以下哪一项行为由数据节点（DataNodes）执行（）

【A.】存储数据块

【B.】创建和删除文件

【C.】维护元数据

【D.】处理客户端请求

【答案】A

与管理

下列各项中，关于HDFS的描述，哪个是错误的（ ）

【A.】支持超大文件存储

【B.】是一个分布式文件系统

【C.】非结构化数据存储效率低

【D.】适应了硬件故障是常态的大数据环境

【答案】C

与管理

下列关于数据仓库三层架构的描述中，正确的是（ ）

【A.】顶层由联机分析处理服务器组成

【B.】底层由数据仓库服务器组成

【C.】中间层由前端用户界面表示

【D.】数据在中间层完成加载和存储

【答案】B

与管理

以下哪个不是NoSQL数据库的类型（）

【A.】关系型数据库

【B.】文档型数据库

【C.】列存数据库

【D.】图数据库

【答案】A

与管理

HDFS架构中的主节点（名称节点）负责（）

【A.】接收来自客户端的作业请求

【B.】接收来自用户的数据读取请求

【C.】创建数据块

【D.】删除数据块

【答案】A

与管理

下列数据库中，属于文档数据库的是（ ）

【A.】Redis

【B.】Neo4j

【C.】HBase

【D.】MongoDB

【答案】D

与管理

下列各项属于数据仓库的特点的是（）

【A.】数据以主题为导向，提供决策支持

【B.】数据源单一

【C.】数据质量低

【D.】不支持历史数据分析

【答案】A

与管理

对于数据垄断问题，一种可能的后果是（）

【A.】使得更多小企业有机会获取数据

【B.】减少企业在市场上的竞争力

【C.】影响其他企业的决策方向并产生依赖

【D.】广告商对大数据平台产生反感

【答案】C

、隐私与合规

在大数据中，以下哪项可能导致个人信息被泄露？

【A.】数据发布的匿名技术

【B.】数据脱敏过于彻底

【C.】不完善的隐私保护政策

【D.】数据线性规划

【答案】C

隐私与合规

下列关于大数据处理中的法规合规性问题，描述正确的是（）

【A.】大数据处理过程中对个人隐私数据的使用无需获得用户同意

【B.】大数据处理过程中对数据的收集、使用、传输都需要符合相关法规和标准

【C.】大数据处理中，只有数据安全需要考虑法规和合规性问题

【D.】大数据处理中，只要数据加密，就可以避免所有的法规和合规性问题

【答案】B

隐私与合规

大数据环境中，用于检测和防止非法用户伪造身份或越权访问的技术是（ ）

【A.】数据脱敏

【B.】访问控制

【C.】数据水印

【D.】数据匿名化

【答案】B

隐私与合规

在大数据环境中，哪项技术可以保证数据的完整性和不可篡改性（）

【A.】数据分割技术

【B.】数据复制技术

【C.】数据压缩技术

【D.】区块链技术

【答案】D

隐私与合规

下列各项属于数据安全风险的是（ ）

【A.】数据的采集风险

【B.】数据的传输风险

【C.】数据的应用风险

【D.】以上均是

【答案】D

隐私与合规

通过数据变形的方式对敏感数据进行处理，从而降低数据敏感程度的数据处理技术叫做（ ）

【A.】数据发布的匿名技术

【B.】社交网络匿名保护技术

【C.】数据水印技术

【D.】数据脱敏技术

【答案】D

隐私与合规

下列属于数据的应用风险的是（）

【A.】个人账号攻击

【B.】存储能力不足

【C.】基础设施差

【D.】访问控制问题

【答案】A

隐私与合规

在大数据的安全与隐私保护中，下列哪一项是用来防止数据在传输过程中被窃取或篡改的（ ）

【A.】数据脱敏

【B.】数据加密

【C.】数据降维

【D.】数据压缩

【答案】B

隐私与合规

在大数据处理中，用以确保数据完整性、防止被非法篡改的安全技术是（ ）

【A.】数据匿名化

【B.】数据脱敏

【C.】数据水印技术

【D.】数据加密

【答案】C

隐私与合规

以下哪一项是大数据处理涉及的隐私保护问题（ ）

【A.】数据量大

【B.】数据复杂度高

【C.】数据的敏感信息泄露

【D.】数据分析速度慢

【答案】C

隐私与合规

下列哪项不属于数据隐私保护技术的类型（ ）

【A.】数据加密

【B.】数据匿名化

【C.】数据挖掘

【D.】数据水印

【答案】C

隐私与合规

下列关于大数据背景下的数据安全问题，哪一项是错误的（）

【A.】大数据架构复杂，数据来源广泛，易引发数据安全问题

【B.】在大数据环境，数据的安全性、完整性和私密性都需要保护

【C.】大数据背景下的数据安全问题只存在于数据传播过程

【D.】大数据带来的挑战是如何在利用数据价值的同时，保证数据的安全、隐私权的保护和合规性

【答案】C

隐私与合规

下列各项属于合规数据的是（ ）

【A.】非法收集隐私信息数据

【B.】取得使用者同意的个人资料数据

【C.】泄露的隐私信息数据

【D.】垄断数据

【答案】B

隐私与合规

2021年9月1日，我国（ ）正式施行

【A.】数据安全法

【B.】网络安全法

【C.】个人信息保护法

【D.】隐私法

【答案】A

隐私与合规

在基于人口统计学的推荐算法中，平台需要根据哪些个人信息数据为用户进行商品推荐？

【A.】用户的姓名与手机号

【B.】用户的职业与学历

【C.】用户的年龄、性别、地区、偏好兴趣等

【D.】用户的网购历史记录与搜索历史

【答案】C

下列哪项是大数据在电子商务个性化服务中的应用体现？

【A.】大数据用于分析销售情况，以提升营销策略

【B.】通过对用户的浏览与购买行为的分析，提供个性化、定制化的推荐服务

【C.】大数据用于实现对市场趋势的精准预测

【D.】大数据用于提升网络安全

【答案】B

下列哪项并非大数据在电商领域的应用？

【A.】客户画像

【B.】仓储管理

【C.】精准营销

【D.】生态保护

【答案】D

以下哪种推荐算法是在用户对各商品的兴趣和偏好的基础上，通过提取商品的关键特征向用户推荐符合这些特征的商品?

【A.】基于规则的推荐算法

【B.】基于人口统计学的推荐算法

【C.】基于内容的推荐算法

【D.】基于项目的过滤推荐算法

【答案】D

下列各项不属于大数据在银行业的应用的是（ ）

【A.】客户分析

【B.】风险管理

【C.】运营优化

【D.】疾病预防与治疗

【答案】D

下列哪个应不属于大数据在医疗领域的应用（ ）

【A.】疾病诊断

【B.】患者管理

【C.】电子病历分析

【D.】动物饲养管理

【答案】D

下列哪项不是大数据在电商领域的主要应用（）

【A.】用户行为分析

【B.】商品推荐

【C.】存货管理

【D.】航天探索

【答案】D

在数据的变换中，下列哪种处理方式能够消除不同变量之间量纲、数量级等属性的差异，使得各数据可以进行综合分析比较？

【A.】离散化

【B.】对数变换

【C.】标准化

【D.】Box-Cox变换

【答案】C

以下哪种插补方法，对于分类型数据使用众数填补，对于数值型数据使用平均数或中位数填补？（）

【A.】回归插补

【B.】均值插补

【C.】直接删除法

【D.】以上都是

【答案】B

在大数据采集和预处理中，以下哪一项不是数据清洗的主要任务（）？

【A.】消除重复数据

【B.】补全缺失数据

【C.】编制数据模型

【D.】校正数据错误

【答案】C

按照预先设定的搜索规则，编写程序自动浏览并下载目标网页信息的数据采集方法称作（）

【A.】系统日志信息数据采集方法

【B.】ETL工具数据采集方法

【C.】网络爬虫数据采集方法

【D.】传感器数据采集方法

【答案】C

使用Python进行网络爬虫获取网页数据时，正确的步骤是（ ）

【A.】发送请求-获取内容-解析内容-保存数据

【B.】保存数据-发送请求-获取内容-解析内容

【C.】获取内容-解析内容-发送请求-保存数据

【D.】解析内容-保存数据-获取内容-发送请求

【答案】A

以下哪个是开源的ETL工具？

【A.】DataStage

【B.】Informatica PowerCenter

【C.】Kettle

【D.】以上均不是

【答案】C

在进行网络爬虫操作中，获取内容使用的Python请求库是（ ）

【A.】Scrapy

【B.】Numpy

【C.】Requests

【D.】Pandas

【答案】C

下列哪个是常用的网络爬虫工具或框架？

【A.】Illustrator

【B.】Photoshop

【C.】Scrapy

【D.】MS Word

【答案】C

关于数据的离散化，下列描述中错误的是（ ）

【A.】数据的离散化是指将数据由分类型变量变成数值型变量

【B.】离散化可以提高大数据处理的效率

【C.】组距分组是常见的离散化方法之一

【D.】离散化可以实现样本量的缩减

【答案】A

下列哪种不是数据预处理中用于数据清洗的技术（ ）

【A.】空值处理

【B.】数据格式转换

【C.】异常值处理

【D.】数据加密

【答案】D

下列哪项是高维数据带来的挑战？（ ）

【A.】变量越多，模型训练时间越短

【B.】变量维度增加，模型复杂度增加

【C.】高维数据所有的变量都重要

【D.】维度越高，预测能力越强

【答案】B

在离散化处理中，当样本等分成4份，那么分位数点分别是（ ）

【A.】10%, 30%, 60%

【B.】15%, 40%, 70%

【C.】25%, 50%, 75%

【D.】20%, 40%, 60%

【答案】C

下列哪项不属于缺失数据处理的步骤？

【A.】识别缺失数据

【B.】分析导致数据缺失的原因

【C.】数据标准化

【D.】处理缺失值

【答案】C

数据的离散化是指将数据由数值型变量变为（ ）

【A.】连续属性

【B.】分类型变量

【C.】实数属性

【D.】整数属性

【答案】B

在大数据分析中，异常值有可能代表的情况是什么？（）

【A.】正常的数据分布

【B.】严重的数据错误

【C.】期待的研究对象，如银行交易中的违约或欺诈行为

【D.】数据收集的全过程

【答案】C

在大数据的预处理阶段，下列哪一种情况不需要进行数据无量纲化的处理（）

【A.】数据特征的量纲不同，如一项特征是长度，另一项特征是电阻

【B.】数据特征的数值跨度差异大

【C.】数据特征的单位相同，数值跨度比较接近

【D.】想要把数据都转化到同一数量级，便于进行比较分析

【答案】C

关于大数据采集，下列哪项（）是不准确的

【A.】数据可以来自各种来源

【B.】大数据采集只关注结构化数据

【C.】大数据采集涉及数据的抓取、清洗和导入等过程

【D.】大数据采集的过程中可能需要解决的问题包括数据的规模、多样性和实时性，以及数据的质量和数据的安全等问题

【答案】B

在大数据预处理中，缺失值的处理方式一般不包括（ ）

【A.】删除含有缺失值的实例

【B.】用众数或者均值等统计方法填补

【C.】忽略缺失值

【D.】用特殊符号表示

【答案】C

下列关于数据预处理的描述中，错误的是（ ）

【A.】数据预处理是数据挖掘过程中很重要的一步

【B.】数据预处理可以提高后续数据分析的准确率

【C.】数据预处理中，无论何时数据都需要降维处理

【D.】数据预处理包括数据清洗、数据集成等步骤

【答案】C

数据的整理是根据分析目的对数据格式、形态和结构进行处理，其中（ ）能够在不损失或损失较少数据本身价值的情况下压缩数据

【A.】数据的聚合

【B.】数据的提取

【C.】数据的连接

【D.】数据的变换

【答案】A

下列关于异常值的描述中，错误的是（）

【A.】可以使用箱线图检测异常值

【B.】当异常值的数量不是很多时，可以直接将含有异常值的观测记录删除

【C.】可以将异常值视为缺失值，按处理缺失值的方法处理异常值

【D.】异常值的存在不属于数据质量问题，不会影响模型的预测能力

【答案】D

关于数据缺失，在数据预处理过程中，下列哪种方法不是处理缺失值常用的方法（）

【A.】删除含有缺失值的观测记录

【B.】用平均值、中位数或众数代替缺失值

【C.】以特定值代替缺失值，如-999或0

【D.】保留缺失值，对数据不做任何处理

【答案】D

以下哪种方法不适用于在线数据集的实时收集（）？

【A.】网络爬虫

【B.】API抓取

【C.】问卷调查

【D.】流数据捕获

【答案】C

常见的离散化方法中，（）的基本思想是将样本等分成若干份，各组所包含的样本个数相同，组限是相应分位数上的分位值

【A.】组距分组

【B.】分位数分组

【C.】秩分组

【D.】均值-标准差分组

【答案】B

大数据预处理中，下列哪一项是用来处理噪声数据的？

【A.】数据清洗

【B.】数据转化

【C.】数据归约

【D.】数据抽样

【答案】A

哪个工具包含了数据采集和数据预处理的环节，可将分散在企业全局的业务系统数据抽取到临时中间层进行清洗、转换、集成？（）

【A.】ETL工具采集

【B.】系统日志信息数据采集方法

【C.】网络爬虫数据采集法

【D.】传感器数据采集方法

【答案】A

为了检查数据中的错误或者异常值，以下哪项方法不常被使用（）

【A.】数据可视化

【B.】统计描述

【C.】异常值检测

【D.】数据解密

【答案】D

数据预处理的主要目的是（）

【A.】确定数据的来源

【B.】改变数据的存储位置

【C.】提高数据的质量和挖掘效率

【D.】增加原始数据的数量

【答案】C

分位数分组的基本思想是将样本（ ），各组所包含的样本个数相同

【A.】等分成若干份

【B.】按属性值大小排序

【C.】随机分组

【D.】按序分组

【答案】A

哪种情况下，可以直接删除含有异常值的记录（）

【A.】数据量很大，异常值相对较少

【B.】数据量小，异常值相对较多

【C.】数据量和异常值数量相当

【D.】数据全为异常值

【答案】A

大数据预处理过程中，不包括（）？

【A.】数据清理

【B.】数据集成

【C.】数据挖掘

【D.】数据变换

【答案】C

在缺失值的处理方法中，（ ）是利用样本观测之间的相关性来插补缺失值

【A.】均值插补

【B.】回归插补

【C.】多重插补

【D.】k近邻算法插补

【答案】D

在大数据处理的过程中，将不同的数据源合并在一起，提供一个统一的数据视图属于（ ）

【A.】数据采集

【B.】数据清洗

【C.】数据集成

【D.】数据选择

【答案】C

数据降维的优势不包括（ ）

【A.】提高预测精度

【B.】降低处理成本

【C.】降低解释性

【D.】减少噪声

【答案】C

在进行大数据预处理中，哪种语言的库为我们提供了大量预处理函数（）

【A.】Python

【B.】C++

【C.】Java

【D.】Go

【答案】A

下列各项不属于大数据采集方法的是（ ）

【A.】系统日志数据采集

【B.】逐一观测采集

【C.】ETL工具采集

【D.】网络爬虫

【答案】B

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.】包含结构化、半结构化和非结构化数据

【答案】C

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

在大数据采集中，以下哪项技术是用于从网页上自动抓取信息的？

【A.】ETL

【B.】网络爬虫

【C.】数据镜像

【D.】数据流转

【答案】B

在处理包含缺失值的数据时，（）的办法可能使得数据分布发生变化

【A.】删除法

【B.】均值插补

【C.】回归插补

【D.】忽略处理

【答案】B

下列哪一项不是常用的数据预处理步骤（ ）

【A.】数据清洗

【B.】数据集成

【C.】数据归约

【D.】数据可视化

【答案】D

数据异常值的处理方法不包括（ ）

【A.】极小值替换

【B.】删除

【C.】忽略

【D.】视为缺失值进行填补

【答案】A

关于数据预处理的描述，以下哪项是不正确的（）

【A.】数据预处理的目的是删除重复的数据

【B.】数据清洗中包括处理缺失值和异常值

【C.】数据离散化可以简化数据结构

【D.】数据变换中包含了数据的平滑、聚集、概化及属性构造

【答案】A

大数据采集过程中，以下哪种工具不能用于数据抓取（ ）

【A.】网络爬虫

【B.】API接口

【C.】数据库导出

【D.】Photoshop

【答案】D

以下哪种方法不适宜处理大数据中的异常值（ ）

【A.】直接删除异常值

【B.】将异常值替换为平均值

【C.】将异常值视为缺失值，用插值法处理

【D.】直接将异常值赋值为极大值

【答案】D

下列哪一项不属于数据预处理的目标（ ）

【A.】提高数据质量

【B.】提高数据安全性

【C.】提高数据的可利用性

【D.】减少数据处理的复杂性

【答案】B

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.】包含结构化、半结构化和非结构化数据

【答案】D

异常值的处理方法包括（ ）

【A.】删除

【B.】视为缺失值进行填补

【C.】忽略

【D.】以上均是

【答案】D

下列关于离散化处理的描述中，错误的是（ ）

【A.】离散化处理可以提高大数据处理的效率

【B.】在调查问卷中，离散化处理可以得到更高的响应率

【C.】数据经过离散化后，可以获得更好的模型解释性

【D.】离散化处理不会影响预测精度

【答案】D

下列各项不属于数据预处理方法的是（ ）

【A.】数据采集

【B.】数据整理

【C.】数据变换

【D.】异常值检测

【答案】A

下列关于缺失值的描述中，错误的是（）

【A.】缺失值是指数据集中有些变量的一个或多个取值无法获得

【B.】数据缺失的现象大量存在

【C.】回归插补的方法不会改变数据分布

【D.】当缺失数据的记录所占比例在数据集中少于10%时，可以将缺失值直接删除

【答案】C

在数据可视化中，适合表示数据随时间变化的趋势的是（ ）

【A.】散点图

【B.】折线图

【C.】柱形图

【D.】饼图

【答案】B

数据可视化中，最适合用来表示类别之间的比较的图形类型是（ ）

【A.】柱状图

【B.】雷达图

【C.】散点图

【D.】面积图

【答案】A

在数据可视化中，用于展示数据分布的常用图形类型是（ ）

【A.】折线图

【B.】饼图

【C.】直方图

【D.】散点图

【答案】C

下列哪种图形适合查看不同类别的数据频数（ ）

【A.】直方图

【B.】散点图

【C.】饼图

【D.】折线图

【答案】A

无法用来反映高维数据之间关系的可视化方法是（）

【A.】相关系数矩阵图

【B.】散点图矩阵

【C.】气泡图

【D.】平行坐标图

【答案】C

散点图中的观测点分布得很分散，没有任何规律，表明两个变量（ ）

【A.】完全线性相关

【B.】线性相关

【C.】非线性相关

【D.】没有相关关系

【答案】D

以下哪种图形适合展示时间序列数据？

【A.】饼形图

【B.】散点图

【C.】直方图

【D.】折线图

【答案】D

可以用来查看数值型变量的分布的可视化方法是（）

【A.】箱线图

【B.】直方图

【C.】小提琴图

【D.】以上方法均可以

【答案】D

以下可视化图形中，（ ）是由数据集合中的最大值、最小值、中位数和两个四分位数绘制而成

【A.】柱形图

【B.】饼图

【C.】箱线图

【D.】直方图

【答案】C

以下常用于时间序列数据可视化的方法是（ ）

【A.】折线图

【B.】直方图

【C.】饼图

【D.】箱线图

【答案】A

可视化一个变量随时间的变化，最合适的绘图方法是（ ）

【A.】柱状图

【B.】折线图

【C.】气泡图

【D.】散点图

【答案】B

如果要对比几个类别的数量分布情况，较为合适的可视化方法是（ ）

【A.】饼图

【B.】折线图

【C.】散点图

【D.】柱状图

【答案】D

气泡图中可以用来展示数据信息的属性包括（ ）

【A.】仅横坐标

【B.】仅横坐标和纵坐标

【C.】仅横坐标、纵坐标和气泡大小

【D.】横坐标、纵坐标、气泡大小和气泡颜色

【答案】D

下列哪种图形更适用于展示某项数据在各个子集中的分布？

【A.】饼图

【B.】箱形图

【C.】散点图

【D.】气泡图

【答案】B

当我们想看出不同分类间的数据分布时，最适合的方法是使用（）

【A.】直方图

【B.】箱线图

【C.】饼图

【D.】散点图

【答案】B

数据可视化的目的是（ ）

【A.】提高数据的质量

【B.】减小数据的体积

【C.】简化数据的处理过程

【D.】增强数据的理解性和洞察力

【答案】D

用以表示变量之间关系的一种可视化方法是（）

【A.】饼图

【B.】条形图

【C.】散点图

【D.】直方图

【答案】C

以下哪种方法不适合用于表示两个变量间的关系？

【A.】散点图

【B.】折线图

【C.】直方图

【D.】气泡图

【答案】C

关于数据可视化的描述中，哪一项是错误的（）

【A.】数据可视化的目的是使数据更易理解

【B.】数据可视化只能用于表示定量数据

【C.】数据可视化可以帮助我们发现数据中的模式

【D.】数据可视化可以帮助我们进行更好的决策

【答案】B

对于只有二分属性的分类变量，最适合使用（）进行数据展示

【A.】饼图

【B.】雷达图

【C.】柱形图

【D.】散点图

【答案】A

如果只是研究两个数值变量之间的关系，最常见的可视化方法是（）

【A.】直方图

【B.】散点图

【C.】饼图

【D.】折线图

【答案】B

当我们在展示时间序列数据时，可以采用以下哪种图表类型（）

【A.】饼图

【B.】气泡图

【C.】折线图

【D.】柱状图

【答案】C

对比一个分类变量的各类别的频数占比，不适合使用（ ）

【A.】柱形图

【B.】条形图

【C.】折线图

【D.】饼图

【答案】C

大数据处理流程中的（ ）步骤是将数据转化为图形，以更直观的方式展示和表达

【A.】数据的采

【B.】数据的存储与管理

【C.】数据的可视化

【D.】数据的分析与挖掘

【答案】C

如果要反映某学生在6个学期中每学期平均成绩的变化情况，采用（ ）可视化方法较为合适

【A.】饼图

【B.】折线图

【C.】散点图

【D.】直方图

【答案】B

下列哪种图形适合表示两个类别型变量之间的关系（ ）

【A.】柱形图

【B.】树形图

【C.】散点图

【D.】饼图

【答案】A

为展示两个数值型变量之间的相关关系，常用的可视化方法是（）

【A.】箱线图

【B.】直方图

【C.】饼图

【D.】散点图

【答案】D

为表示一组数据的分布特征，反映数据分布是否对称时，常用的可视化方法是（）

【A.】箱线图

【B.】气泡图

【C.】折线图

【D.】散点图

【答案】A

箱线图在数据分析中的用途，除了表示一组数据的分布特征，还可以反映（ ）

【A.】数据的均值

【B.】数据的标准差

【C.】数据分布是否存在离群点

【D.】数据的趋势

【答案】C

定义并简述NoSQL数据库的特点

【答案】NoSQL数据库，即Not only SQL，意为不仅仅是SQL。它的特点是可以处理大数据量、高并发、分布式的数据库。

请详述大数据的存储与管理技术的发展过程及其演化的原因。

【答案】数据存储与管理技术从文件管理系统开始，后来数据库和数据仓库出现并逐渐占据主导地位，到了大数据时代，随着数据类型的愈发复杂，数据量的急剧增大，能够处理数据的存储和查询技术迅速发展，涌现了如分布式文件系统、NoSQL非关系型数据库等新型的数据管理技术。

请简述大数据的价值所在？

【答案】大数据的价值在于可以通过对海量数据的分析和挖掘，帮助企业发现潜在的商业机会、改善决策过程、提升运营效率、优化用户体验等，从而实现商业增长和竞争优势。

变量的定义是什么？

【答案】用于刻画观测数据集的特征的量叫做变量。

根据数据的结构化程度，可以将数据分为哪三种类型？

【答案】根据数据的结构化程度，可以将数据分为结构化数据、非结构化数据和半结构化数据。

什么是结构化数据？举例说明。

【答案】结构化数据是指存在于一定的数据模型或结构中，可通过预先设计的模型进行处理和分析的数据，如数据库中的表格数据，CSV文件等。

为什么数据的可视化在大数据处理中被视为重要的步骤？它的主要作用是什么？

【答案】数据的可视化将数据转化为图形，以更直观的方式展示和表达数据，可帮助我们从数据中得到有价值的推论或发现某种规律和特征，甚至可以洞察统计分析中无法发现的结构和细节。同时，数据可视化可以在数据分析之前作为数据探索的步骤，也可以出现在数据分析之后用来呈现分析结果。

请简述大数据时代下，数据的变化对现代社会有什么影响？

【答案】大数据时代下，数据的变化对现代社会产生了许多影响。首先，数据的爆炸式增长使得人们可以从更多维度了解和分析现实世界，有助于科学研究和决策制定。其次，数据的共享和开放促进了创新和发展，例如数据驱动的智能系统和服务的发展。此外，数据的挖掘和分析可以帮助企业和组织发现商业机会，提高效率和竞争力。然而，大数据时代也面临着数据隐私和安全的挑战，因此保护个人隐私和数据安全成为一个重要议题。

请解释在数据的存储和管理中，数据库技术是如何发挥作用的？

【答案】数据库技术是一种数据管理的方法和技术，可以降低数据管理的复杂性，使数据的组织和维护变得更方便，从而可以有效地利用数据。随着数据库技术的不断进步，我们已经能够从传统的关系型数据库发展到高性能的非关系型数据库，数据的存储和管理能力得到了不断的提升。具有海量数据存储能力的数据库可以很好地应用于大数据时代的各种类型的数据管理中，包括结构化数据和各种非结构化数据。

非结构化数据在大数据时代的特点和挑战是什么？

【答案】非结构化数据通常包含办公文档、图像、声音、树结构和文本等，不存在或者很难发现统一的数据结构，不能以二维表的形式存储在数据库中。非结构数据的容量大、产生速度快并且来源多样，且非结构化数据增长速度比结构化数据增长速度高。由于没有规定的结构，需要特定软件才能浏览，因此无法直接在传统的关系数据库中存储、管理和处理，给数据分析带来挑战。

大数据在应用中主要面临哪些挑战？

【答案】大数据在应用中主要面临数据的获取、存储、处理和分析的困难，以及数据的安全性和隐私保护的问题。此外，大数据应用还需要解决数据的质量、可靠性、实时性、分布式计算和算法等方面的挑战。

请分别解释结构化数据、非结构化数据和半结构化数据的含义，并各自举一个例子？

【答案】结构化数据是按照固定格式和规则组织的数据，具有明确的数据类型和内部关系。例如，关系型数据库中的表格和字段就是结构化数据的典型例子。非结构化数据是没有固定格式和规则的数据，无法通过简单的表格或字段来表示。例如，电子邮件、图片、音频和视频等都属于非结构化数据。半结构化数据介于结构化数据和非结构化数据之间，它有一定的结构，但不符合严格的模式或规则。例如，XML和JSON格式的数据就是半结构化数据的常见形式。

简述大数据的利用可以带来哪些价值？

【答案】大数据的利用可以带来商业决策优化、新的商业模式、消费者需求预测、市场趋势分析等多方面价值。

请描述定性变量和定量变量的基本定义和区别？

【答案】定性变量是指用于描述和区分对象特征的变量，其取值为非数字类别，如性别、颜色等。定量变量是指用于度量对象特征的变量，其取值为数字，如身高、年龄等。定性变量和定量变量的区别在于其取值类型不同，定性变量是离散的非数字类别，而定量变量是连续或离散的数字。

定量变量的定义是什么？

【答案】当一个变量的取值可以在一个范围内连续取值时，该变量就是定量变量。

请解释什么是截面数据，并给出一个相关的例子。

【答案】截面数据是指同一时间不同观测个体的一个或多个指标组成的数据集，例如2022年我国各省的各项经济指标数据集就是截面数据。

数据的可视化在大数据处理流程中的作用是什么？给出一个例子。

【答案】数据的可视化在大数据处理流程中的作用是将复杂的数据以图表、图像等可视化形式展示，使得数据更容易理解和分析。例如，通过绘制柱状图或折线图展示销售数据，可以直观地观察到销售额的变化趋势和销售量的分布情况。

什么是数据的可视化？

【答案】数据的可视化是指通过图表、图形等视觉元素将数据或数据间关系展示出来，使人更直观、更简单地理解数据和数据间的关系。

大数据的成因是什么？

【答案】数据的存储和管理能力的增强；数据采集能力增强；大数据的挖掘和分析等技术的同步发展。

Spark和Hadoop在处理大数据时的主要区别是什么？

【答案】Spark是内存计算工具，对于需要反复迭代的数据处理有很大处理速度优势；Hadoop的MapReduce更适合大批量的数据处理，处理效率稳定。

数据的定义是什么？

【答案】数据是对现象或事物进行测量和记录的结果，可用来制表、计算和分析等，也可以统指一切保存在电脑中的信息，能够进行电子化的记录，包括文本、图像、音频、视频等。

大数据的处理流程主要包括哪5个步骤？

【答案】数据的采；数据的存储与管理；数据的可视化；数据的分析与挖掘；大数据的处理。

为什么我们在做聚类分析之前需要进行特征选择或者特征抽取？

【答案】因为特征并不是越多聚类效果就越好，有些特征中包括噪声，对聚类结果没有贡献反而降低了聚类效果。所以，可以使用一些最有代表性的特征或者将原特征进行组合变换构成新的特征。

简述协同过滤推荐算法的基本原理及其优缺点？

【答案】协同过滤推荐算法是一种基于用户行为的推荐方法，基本原理是通过分析用户的历史行为、兴趣和偏好，找到和该用户类似的其他用户，然后将这些用户喜欢的物品推荐给该用户。优点是能够发现用户之间的共同兴趣，推荐准确度较高；缺点是对新用户和冷启动问题处理不好，需要大量的用户行为数据才能进行推荐。

通过相关系数矩阵处理共线性问题的算法步骤是什么？

【答案】计算相关系数矩阵；确定最大的成对相关系数对应的预测变量A和预测变量B；计算变量A与其他所有变量之间的平均绝对值相关系数，对变量B也做同样的计算；比较A与B，谁的平均绝对值相关系数最大，删除谁；重复以上步骤，直到两两之间绝对值相关系数低于某一特定阈值。

请简述基于人口统计的推荐算法有何特点及应。

【答案】基于人口统计的推荐算法主要是根据用户的人口学信息（如性别，年龄，职业等）给用户推荐物品，其特点是不需要用户的历史行为数据，只需要用户的基本信息，适包括新用户推荐等场合。

请简述使用kNN算法将未分类的观测或个体划分到跟它最接近的观测所在类别的具体实施步骤。

【答案】1）选择近邻的数量k和距离度量函数；2）找到与未知样本距离最近的k个训练样本；3）看最相似的k个样本多数是哪一类，就把未知样本划分为哪一类。

请简述决策树在大数据分析中的应用以及其主要优缺点？

【答案】决策树在大数据分析中具有广泛应用，可以用于分类和回归问题。它的主要优点包括易于理解和解释、能处理多种数据类型、对缺失值不敏感等。然而，决策树也存在一些缺点，如容易过拟合、对异常值敏感等。

在具体实施时，聚类分析的四个步骤分别是什么？

【答案】（1）特征选择或特征抽取 （2）聚类算法的设计与选择 （3）聚类的效果评估 （4）聚类的解释

请简述什么是协同过滤推荐算法以及其基本工作原理。

【答案】协同过滤推荐算法是一种通过收集多用户的偏好信息来预测当前用户可能感兴趣的信息的推荐算法，它的基本原理主要是基于用户的行为（例如购买或评分）来找出相似的用户或物品，然后进行推荐。

请解释K均值聚类算法的基本步骤是什么？

【答案】K均值聚类算法的基本步骤有：① 先随机取K个点作为初始的聚类中心；② 计算数据中每个点到K个聚类中心的距离（例如欧氏距离），距哪个类中心点更近就将这一样本点归入哪个类别之中；③ 所有样本归类完成以后，重新计算每一类中所有样本点的中心点，这些点确定为新的聚类中心，重复迭代上述过程，直至类中心不再变化为止。

请解释什么是分类方法，包括训练集和检验集，并给出一些常见的分类算法。

【答案】分类方法是根据输入数据建立属性集X和类标号Y之间联系的分类模型。训练集用于建立分类模型，类别标签已知，而检验集中的类别标签未知，用于预测其类别。常见的分类算法包括k近邻法，朴素贝叶斯分类器，C5.0，AdaBoost，随机森林，支持向量机等。

请列举三种常用的电子商务推荐算法。

【答案】协同过滤推荐算法；基于内容的推荐算法；基于关联规则的推荐算法；基于人口统计的推荐算法。

请列举三种大数据在零售业的应用。

【答案】库存优化（如预测商品销售量）、客户行为分析、智能推荐等。

请列举三种大数据在金融领域的应用。

【答案】客户分析（如个人客户画像）、风险管理（如中小企业贷款风险评估）、运营优化（如产品优化）等。

请列举三种大数据在教育领域的应用。

【答案】个性化教学（如根据学生学习状况制定个性化教学计划）、学生成绩预测、课程优化等。

银行如何利用大数据进行风险管理？

【答案】银行使用大数据技术可以有效分析经营过程中可能会存在的风险因素，比如通过整合银行内外部数据，了解客户贷前准入情况、贷后偿债风险等信息，及时发现并控制客户风险。

请解释一下车联网是什么，它在实现智能化交通管理、智能动态信息服务和车辆控制上有何重要意义？

【答案】车联网是由车辆位置、速度和路线等信息构成的巨大交互网络，对实现智能化交通管理、智能动态信息服务和车辆控制具有重要的意义，如，帮助实现交通拥堵的预测和控制，提供及时准确的路况信息，提升驾驶的安全性等。

请描述无人驾驶汽车如何依靠车联网技术提供的数据实现自动行车。

【答案】无人驾驶汽车主要依靠的是车内以计算机系统为主的智能驾驶系统，通过车载传感器感知周边道路环境，车联网在环境感知、计算决策和控制执行等环节上为无人感知提供数据支持。自动规划行车路线，时刻控制车辆转向、速度等参数，保证车辆时刻在规定计划内正常行驶，从而达到预定目标。

在教育领域，大数据主要能够带来哪些创新应用？

【答案】大数据在教育领域可以带来许多创新应用。首先，大数据可以用于学生学习行为的分析，从而提供个性化的学习推荐和指导。其次，大数据可以用于教学内容和教学方法的优化，帮助教师更好地设计和实施教学。此外，大数据还可以用于教育评估和学生综合素质评价，提供更全面、客观的评价结果。最后，大数据还可以用于教育决策和政策制定，为教育改革提供科学依据。

请列举三种大数据在广告行业的应用。

【答案】精准推广（如根据客户画像推送广告）、广告效果评估、竞品分析等。

请列举三种大数据在汽车领域的应用。

【答案】精准营销、交通事故预防和无人驾驶等。

为什么要进行数据的提取操作？其在数据整理中起到了什么作用？

【答案】在大数据时代，由于原始数据量大，直接处理所有数据的情况是非常少见的。通常我们的分析目标只需要部分数据，所以我们需要进行数据的提取。提取可以找到一个适合内存处理并且可以回答问题的变量子集、子样本或者摘要数据。通过提取，可以有效地降低数据量，降低计算成本，并提高计算效率。

为什么需要对高维数据进行降维？请举例说明它在现实生活中的应用。

【答案】随着数据采集技术的发展，我们获取的数据维度越来越高，这将导致数据处理的复杂性增加。通过降维，我们可以消除冗余的信息，增加计算的效率，同时也能减轻一些学习算法的“维度诅咒”问题。例如，在生物医学领域，我们可以通过降维找到对疾病发展有重要影响的关键基因。

请说明什么是数据降维，在大数据处理中，其作用是什么？

【答案】数据降维是指通过保留原始数据中最具代表性的特征，减少数据维度的过程。在大数据处理中，数据降维的作用是减少数据的存储空间和计算复杂度，提高数据处理的效率。

什么是数据转换在大数据预处理中的作用，提供一个实际应用的例子？

【答案】数据转换是指对原始数据进行格式转换、属性变换、值变换等操作，以满足分析和挖掘的需求。它在大数据预处理中起到重要作用，可以通过将原始数据转换为可用于分析的形式，提高数据的可理解性和可用性。例如，在进行情感分析时，可以将原始的文本数据转换为文本向量表示，以便进行情感判别和情感趋势分析。

什么是缺失值？如何处理缺失值？

【答案】缺失值是指在数据集中某一变量的一个或多个观测值未被记录或无法获得的情况。缺失值处理方法众多，常见的方法包括删除缺失值、数据插补或不进行处理。

当预测变量包含的缺失值比率较高时，我们一般使用什么方法处理？为什么要这样做？

【答案】当预测变量包含的缺失值比率较高时，我们通常选择将其删除。因为这种变量包含有用信息的可能性比较低，删除后对模型的性能影响不大。

什么是多重插补方法，与其他插补方法相比有何特点和优势？

【答案】多重插补方法是基于重复取样和插补的思想进行数据缺失值处理的一种方法。它通过多次复制原始数据集，并在每个复制数据集中进行插补，最后将多个插补结果进行合并，得到一个最终的完整数据集。与其他插补方法相比，多重插补方法能够更好地利用缺失值之间的关联关系，提供更准确和可靠的插补结果。它同时还能够生成多个插补模型，可以通过模型之间的比较和分析来评估不确定性，更好地反映了数据的不确定性。因此，多重插补方法在处理数据缺失值时具有较高的灵活性和稳健性，广泛应用于研究和实践中。

简述大数据预处理中的数据清洗是什么，并举例说明其重要性？

【答案】数据清洗是指对原始数据进行去除错误、缺失、冗余、重复等无效或不符合要求的数据处理操作。它在大数据预处理中非常重要，可以保证数据的质量和可用性，提高后续分析和挖掘的准确性和可信度。例如，在进行用户行为分析时，如果原始数据中包含重复记录或错误数据，可能导致分析结果产生偏差，影响决策和业务效果。

请简述解释大数据中的“脏数据”是什么以及可能的产生原因。

【答案】“脏数据”是指存有错误或者不一致性的数据，可能的产生原因包括数据采集过程中的误差、数据转换过程中的不一致、系统性的数据错误等。

数据离散化的定义是什么？

【答案】数据的离散化是指将数据由数值型变量变成分类型变量，即将变量的取值由原来的一个区间内连续取值映射到为若干个有限个值。

请简述数据离散化的主要作用以及常用的数据离散化方法。

【答案】数据离散化可以降低数据复杂度，便于分析处理。常用的数据离散化方法有等宽法、等频法和基于聚类分析的离散化方法等。

请简述异常值检测的作用以及常用的异常值检测方法。

【答案】异常值检测的作用主要是找出与正常数据显著不同的数据点，其常用方法包括基于统计的方法（如3σ法），基于距离的方法（如k近邻法）和基于密度的方法等。

请列举五种常见的数据缺失值插补方法。

【答案】均值插补；回归插补；随机回归插补；多重插补；K近邻算法插补。

数据整理的内容主要包括哪四个方面？

【答案】数据的提取；数据的连接；数据的聚合；去除冗余和重复。

什么是数据探索性分析？

【答案】数据的探索性分析是利用图形和统计方法来揭示数据背后的模式、结构、关系，判断数据的分布和异常值等。

饼图适用于哪些类型的数据可视化？

【答案】饼图适用于表示各类别数据占总体数据的比例，便于比较各类别之间的相对大小。

数据可视化的作用是什么？

【答案】快速获取信息；数据的探索性分析；挖掘数据隐藏的规律。

什么是数据的探索性分析，使用数据可视化有哪些优势？

【答案】数据的探索性分析是指通过对数据进行统计分析和可视化展示，发现数据中存在的模式、趋势、异常等信息。使用数据可视化进行探索性分析有以下优势：能够直观地展示数据的分布、关系和趋势；能够帮助快速发现数据中的异常值或离群点；能够辅助决策和沟通，使复杂的数据更易理解和传达。

数据可视化的基本原则包括哪些方面？

【答案】数据可视化的效果要能准确的表达数据中的信息而不产生偏差或歧义；能够清晰地表达数据中的信息；其设计的可视化图表能够令人赏心悦目。

为什么说数据可视化能增强数据的可理解性和易用性？

【答案】因为数据可视化将数据转化为图形和图像，使得数据交流更直观，便于人脑理解和记忆，从而增加数据的可理解性和易用性。

请简述散点图在数据可视化中的应用及其主要优点？

【答案】散点图在数据可视化中广泛应用。主要优点有：1.可以展示两个变量之间的关系，帮助发现变量之间的模式和趋势；2.可以发现异常值和离群点，帮助识别数据的特殊情况；3.可以用不同的颜色、大小、形状等属性来表示其他变量的信息，增加了可视化的维度；4.可以进行多变量的比较和分析，帮助理解多个变量之间的交互作用。

数据分布可视化中，条形图与直方图的区别是什么？

【答案】条形统计图中，横轴上的数据是孤立的，是一个具体的数据，而直方图中，横轴上的数据是连续的，是一个范围；条形统计图是用条形的高度表示频数的大小，而直方图是用长方形的面积表示频数；条形统计图中，各个数据之间是相对独立的，各个条形之间是有空隙的，而在直方图中，各长方形对应的是一个范围，而在直方图中，长方形之间没有空隙。

在选择数据可视化工具时，需要考虑哪些因素？请举例说明。

【答案】选择数据可视化工具时需要考虑以下因素：数据类型和特点，例如选择适合处理时序数据的工具；可视化需求和目标，例如是否需要实时更新、交互性等；使用的平台和环境，例如需要支持移动设备还是只在桌面使用；个人技能和经验，例如是否熟悉某个具体工具的使用。举例说明，如果要可视化大规模地理空间数据，可以选择使用ArcGIS等GIS工具；如果需要实时监控和分析数据，可以选择使用Kibana等实时可视化工具。

请简述数据可视化的基本概念以及其在数据分析中的主要作用。

【答案】数据可视化是将数据转换成可视化的图表，以更直观的方式展现或呈现数据，增强人脑对信息的理解。它在数据分析中的主要作用是提高信息分析效率，揭示数据背后的价值，发现规律，以及预测未来。