

重庆数字资源集团数据运营分公司

2026 年招聘考试押题试卷

全卷共 61 题，考试时间 120 分钟

一、单选题（共 51 题） 本部分共 51 题。

1. 某数据运营团队计划对一批数据进行清洗，原计划每天处理 1200 条，实际每天比原计划多处理 20%，结果提前 2 天完成任务。则这批数据共有多少条？
A. 14400
B. 16800
C. 19200
D. 21600
2. 在数字经济时代，数据已成为新的生产要素。对于数据运营岗位而言，理解数据资产的战略价值至关重要。下列对数据资产特征的描述，最准确的一项是：
A. 数据资产具有排他性，一旦被某人使用，他人便无法使用
B. 数据资产的价值会随着使用次数的增加而递减
C. 数据资产具有非竞争性和部分排他性，其价值在于开发利用
D. 数据资产与实物资产相同，均遵循边际效用递减规律
3. 数据要素显化是指将原本隐藏在业务流程中的非结构化数据，通过技术手段转化为可量化、可价值评估的结构化数据的过程。根据上述定义，下列属于数据要素显化的是：
A. 某银行通过日志分析，将用户操作行为转化为风险评分指标
B. 某企业将纸质合同扫描成电子档案存储
C. 某政府机构在门户网站公布年度财政预算报告
D. 某学校统计在校学生人数并上报教育局
4. 某数据中心的服务器按照每小时 15% 的速度增长数据量。若初始数据量为 100GB，则经过 3 小时后数据量约为多少 GB？（结果保留整数）
A. 152
B. 157
C. 162
D. 167
5. 在数据运营工作中，对数据质量的评估通常包括多个维度。下列选项中，最能准确概括数据质量核心要求的是：
A. 完整性、准确性、一致性、及时性

- B. 多样性、海量性、高速性、价值性
 - C. 可采集、可存储、可计算、可展示
 - D. 标准化、流程化、自动化、智能化
6. ”数据：清洗”与下列哪组词语关系最为类似？
- A. 文件：归档
 - B. 任务：完成
 - C. 食材：烹饪
 - D. 图书：阅读
7. 某数据标注团队有 10 名员工，其中 6 人会使用 Python，5 人会使用 R 语言，3 人两种都会。现从中随机选取 2 人，则这 2 人至少有一人会使用 Python 的概率是多少？
- A. $\frac{2}{3}$
 - B. $\frac{3}{5}$
 - C. $\frac{8}{15}$
 - D. $\frac{13}{15}$
8. 某数据运营团队需要从海量用户行为日志中提取关键指标，以下哪种工具组合最适合用于分布式数据存储和批处理计算？
- A. HDFS + MapReduce
 - B. MySQL + Redis
 - C. MongoDB + Elasticsearch
 - D. SQLite + Excel
9. 某研究机构对多家企业调研后发现，大数据企业的“算法优化”投入与营业收入增长率呈显著正相关，因此研究人员认为，企业应加大算法优化投入以提升营业收入。以下哪项如果为真，最能削弱上述结论？
- A. 该调研样本集中于互联网行业，未涉及制造业
 - B. 多数企业在营收增长后才开始增加算法优化投入
 - C. 部分算法优化项目因技术难度大而失败
 - D. 算法优化投入占企业总投入的比例平均不到 10%
10. 数列：2，3，5，9，17，？
- A. 31
 - B. 33
 - C. 35
 - D. 37
11. 根据《中华人民共和国数据安全法》，数据处理活动应当遵循的原则不包括：

- A. 数据安全与数字经济发展并重
 - B. 以数据开发利用优先于安全保护
 - C. 采取必要措施保障数据安全
 - D. 建立健全数据安全治理体系
12. 数据经纪人是指经数据提供方或数据需求方授权，在两者之间提供数据产品撮合、数据质量评估、交易合规审查等服务的专业机构或个人。根据上述定义，下列不属于数据经纪人行为的是：
- A. 某公司受数据供应商委托，对拟出售的数据产品进行脱敏合规审查，并协助对接需求方
 - B. 某平台基于用户浏览记录向用户推荐相关商品信息
 - C. 某评估机构接受买方委托，对卖方的数据资产进行质量评级并出具评估报告
 - D. 某中介人为数据买卖双方提供交易价格协商和合同签订服务
13. 某数据产品定价为每份 50 元，每月可售出 2000 份。若每份降价 1 元，月销量增加 100 份。问定价为多少元时，月收入最大？
- A. 35
 - B. 40
 - C. 45
 - D. 50
14. 在 SQL 中，用于将多个查询结果合并为一个结果集且去除重复行的关键字是：
- A. UNION
 - B. UNION ALL
 - C. JOIN
 - D. INTERSECT
15. 某数据管理系统的规则如下：所有通过数据质量检测的数据，都完成了数据清洗；所有完成数据清洗的数据，均通过了格式转换。由此可以推出：
- A. 所有通过格式转换的数据都通过了数据质量检测
 - B. 所有通过数据质量检测的数据都通过了格式转换
 - C. 有些未通过格式转换的数据通过了数据质量检测
 - D. 有些未通过数据质量检测的数据未通过格式转换
16. 在数据治理体系中，用于描述数据定义、来源、质量、血缘关系及使用约束等属性的“数据”通常被称为：
- A. 主数据，描述企业核心业务实体及其属性
 - B. 元数据，描述数据定义、来源、血缘与约束
 - C. 参考数据，用于业务分类和编码的标准数据
 - D. 衍生数据，由原始数据处理后得到的数据

17. 某数据运营团队计划在 5 个工作日内完成一项数据清洗任务，前 2 天每天完成总量的 15%，第 3 天完成总量的 20%，后 2 天平均每天完成剩余部分的多少才能按时完成任务？
- A. 22.5%
 - B. 25%
 - C. 27.5%
 - D. 30%
18. 某数据运营团队使用 Spark 进行数据清洗，发现某列数值存在大量异常值，以下哪种方法最适合处理该情况？
- A. 删除包含异常值的所有行
 - B. 用均值或中位数替换异常值
 - C. 忽略异常值，直接进行后续分析
 - D. 将异常值全部转换为空值
19. ”谨慎对于冒进”相当于：
- A. 合规对于违规
 - B. 创新对于改革
 - C. 效率对于质量
 - D. 安全对于风险
20. 关于 Hadoop 生态组件，下列说法正确的是：
- A. HDFS 适合存储大量小文件，NameNode 压力较小
 - B. MapReduce 基于内存迭代计算，适合实时流处理
 - C. YARN 负责集群资源管理和任务调度
 - D. Hive 是分布式内存计算引擎，用于交互式查询
21. 某数据产品定价为每份 120 元，成本为 80 元。若每份降价 5 元，成本不变，则降价后每份的利润率比降价前下降了多少个百分点？
- A. 5.0
 - B. 6.25
 - C. 7.5
 - D. 8.0
22. 在数据生命周期管理中，数据销毁环节的主要目的是：
- A. 释放存储空间
 - B. 防止数据泄露
 - C. 提高数据质量
 - D. 降低存储成本

23. 某数据运营团队的四位员工关于“谁使用了数据库管理权限”有如下陈述：甲：我没使用。乙：丙使用了。丙：丁使用了。丁：丙没有使用。已知四人中只有一人说了真话，由此可以推出：
- A. 甲使用了权限
 - B. 乙使用了权限
 - C. 丙使用了权限
 - D. 丁使用了权限
24. 在数据采集与预处理中，使用 Z-score 标准化处理连续型特征的主要目的是：
- A. 消除量纲影响，使特征均值为 0、标准差为 1
 - B. 将特征值映射到 [0,1] 区间，便于统一比较
 - C. 通过取对数压缩数据分布，减少异常值的影响
 - D. 将分类变量转换为数值型变量，便于模型输入
25. 某数据中心的服务器每天处理的数据量呈线性增长，第 1 天处理 10TB，第 5 天处理 30TB。若按此趋势，第 10 天处理的数据量是多少？
- A. 55TB
 - B. 60TB
 - C. 65TB
 - D. 70TB
26. 某数据运营团队需要从多个数据源采集数据，以下哪种工具最适合实时数据流处理？
- A. Apache Kafka
 - B. Sqoop
 - C. Flume
 - D. DataX
27. 某调查机构对数据运营岗位员工进行为期两年的跟踪研究发现，定期参加数据技能培训的员工，其绩效评分平均比未参加培训的员工高 15%。因此，该机构建议企业增加数据技能培训投入。以下哪项如果为真，最能支持上述建议？
- A. 培训课程内容与员工实际工作场景高度相关
 - B. 未参加培训的员工大多有五年以上工作经验
 - C. 参加培训的员工在数据工具使用熟练度上显著提升
 - D. 部分绩效评分高的员工也主动参加了培训
28. 在关联规则挖掘中，给定事务数据集，规则 $A \rightarrow B$ 的置信度（confidence）定义为：
- A. $\text{support}(A \cup B) / \text{support}(A)$
 - B. $\text{support}(A \cup B) / \text{support}(B)$
 - C. $\text{support}(A \cup B) / N$

D. $\text{support}(A) / \text{support}(B)$

29. 某数据产品上线后，日活跃用户数从 1000 人增长到 2500 人，用了 3 个月。若保持相同的月增长率，再过 3 个月日活跃用户数约为多少人？
- A. 6250
B. 6400
C. 6550
D. 6700
30. 在数据挖掘中，以下哪种算法属于无监督学习？
- A. 线性回归
B. 决策树
C. K-means 聚类
D. 支持向量机
31. 某数据公司从重庆、成都、贵阳、昆明四地招聘了甲、乙、丙、丁四名数据人才，四人分别负责数据运营、数据治理、数据安全、数据开发四份工作。已知：（1）来自重庆的员工不负责数据运营；（2）来自成都的员工负责数据安全；（3）甲来自重庆，且不负责数据开发；（4）乙不来自贵阳，也不负责数据运营；（5）丙负责数据治理；（6）丁不负责数据开发。根据以上信息，以下哪项一定为真？
- A. 甲负责数据安全
B. 乙来自昆明
C. 丙来自成都
D. 丁负责数据开发
32. 根据《中华人民共和国个人信息保护法》，处理敏感个人信息应当取得个人的（ ），并将处理敏感个人信息的必要性以及对个人权益的影响告知个人。
- A. 默示同意
B. 单独同意
C. 概括同意
D. 书面同意
33. 某数据运营公司计划投资一个项目，初始投资 100 万元，预计每年末产生收益 30 万元，持续 5 年。若年贴现率为 10%，则该项目净现值为多少万元？
- A. 13.72
B. 15.86
C. 17.24
D. 19.08
34. 根据《个人信息保护法》，处理个人信息应当遵循的原则不包括：

- A. 合法、正当、必要
 - B. 公开、透明
 - C. 目的限制
 - D. 信息自由流通优先
35. 某市大数据发展报告显示，2025 年该市数字经济核心产业增加值占 GDP 比重为 12.5%，较 2020 年提高了 3.2 个百分点；“十四五”期间，该市数字经济核心产业增加值年均增速超过 15%。根据上述信息，以下哪项一定为真？
- A. 2020 年该市数字经济核心产业增加值占 GDP 比重为 15.7%
 - B. 2025 年该市 GDP 比 2020 年下降了
 - C. 2020 年该市数字经济核心产业增加值占 GDP 比重为 9.3%
 - D. “十四五”期间该市 GDP 年均增速超过 15%
36. 根据财政部《企业数据资源相关会计处理暂行规定》（财会〔2023〕11 号），企业将数据资源按照存货准则进行会计处理时，应满足的前提是：
- A. 企业在日常活动中持有，最终目的是用于出售的数据资源
 - B. 企业用于内部研发、管理和运营的数据资源
 - C. 企业通过数据加工形成并自用的数据产品
 - D. 企业从公开渠道免费获取且无法单独辨认的数据
37. 某数据运营团队需要从 6 个项目中选 3 个进行优先实施，其中项目 A 和 B 必须至少选一个，则共有多少种选法？
- A. 16
 - B. 18
 - C. 20
 - D. 22
38. 在数据运营中，用于衡量用户留存情况的指标是：
- A. 日活跃用户数（DAU）
 - B. 月活跃用户数（MAU）
 - C. 用户留存率
 - D. 用户获取成本
39. 某电商平台在“双十一”期间点击量和交易量激增，但后台监控显示，其数据实时计算集群的 CPU 平均利用率反而低于平时。以下哪项如果为真，最能解释这一现象？
- A. 该平台在流量高峰前自动扩展了计算节点数量，负载被分散到更多节点
 - B. “双十一”期间平台暂停了部分非核心业务系统，减少了计算任务
 - C. 平台在平时有大量的离线计算任务，拉高了 CPU 利用率

- D. “双十一”期间用户访问路径变短，单次请求的 CPU 消耗降低
40. 关于 Apache Spark 与 Hadoop MapReduce 的对比，下列说法正确的是：
- A. Spark 基于内存计算，在迭代式计算场景下通常比 MapReduce 性能更高
 - B. Spark 只能处理流式数据，无法处理批量数据
 - C. MapReduce 采用 DAG 调度，Spark 采用简单的两阶段作业调度
 - D. Spark 必须依赖 MapReduce 才能读取 HDFS 数据
41. 某数据运营团队需要对一批数据进行清洗，原计划每天处理 1200 条，实际每天比计划多处理 20%，结果提前 2 天完成任务。这批数据共有多少条？
- A. 12000 条
 - B. 14400 条
 - C. 16800 条
 - D. 19200 条
42. 某数据运营团队使用 Hive 进行数据仓库建设，以下哪种文件格式最适合存储高压缩比且支持列式查询的数据？
- A. 文本文件（TEXTFILE）
 - B. 序列文件（SEQUENCEFILE）
 - C. Parquet
 - D. CSV
43. 某数据公司提出，将数据质量考核与员工绩效奖金挂钩，可以有效提高公司的数据质量水平。为使上述主张成立，需要补充以下哪项作为前提？
- A. 员工数据质量考核指标不透明，难以量化
 - B. 员工有能力和权限改善其所负责的数据质量问题
 - C. 数据质量水平主要依赖底层技术架构的稳定性
 - D. 绩效奖金挂钩后，部分员工可能虚报数据质量指标
44. 关于数据质量维度，下列说法正确的是：
- A. 准确性是指数据必须符合预先定义的值域、格式和业务规则
 - B. 完整性是指数据应真实反映客观实体及其业务活动
 - C. 一致性是指同一逻辑实体在不同数据源中记录的值保持一致
 - D. 及时性是指数据从产生到可被使用的时间间隔越短越好
45. 某数据产品上线后，日活跃用户数（DAU）从第 1 天的 1000 人开始，之后每天比前一天增长 10%，但第 5 天因系统升级导致 DAU 比前一天下降 5%。请问第 5 天的 DAU 是多少人？
- A. 1300 人

- B. 1331 人
 - C. 1264.45 人
 - D. 1250 人
46. 在数据可视化中，以下哪种图表最适合展示数据随时间变化的趋势？
- A. 饼图
 - B. 折线图
 - C. 散点图
 - D. 雷达图
47. 某数据分析师发现，过去一年中，公司每逢推出促销活动，当月销售额都会上升。因此他认为，促销活动是销售额上升的原因。以下哪项最准确地指出了上述论证的缺陷？
- A. 促销活动与销售额上升之间不具有统计学相关性
 - B. 将时间上的先后关系错误地当成了因果关系
 - C. 没有区分不同促销活动的具体类型
 - D. 忽略了销售额上升可能是由其他因素单独导致的
48. 在数据生命周期管理框架中，下列各阶段排序最合理的是：
- A. 采集 → 存储 → 处理 → 分析 → 共享 → 归档 → 销毁
 - B. 采集 → 处理 → 存储 → 分析 → 共享 → 归档 → 销毁
 - C. 存储 → 采集 → 处理 → 分析 → 共享 → 销毁 → 归档
 - D. 采集 → 存储 → 分析 → 处理 → 共享 → 销毁 → 归档
49. 根据《数据安全法》，国家建立数据分类分级保护制度，下列哪项属于核心数据？
- A. 一般工业数据
 - B. 公共数据
 - C. 关系国家安全、国民经济命脉的重要数据
 - D. 个人一般信息
50. 根据《中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见》（“数据二十条”），关于数据产权制度框架，下列说法正确的是：
- A. 建立数据资源持有权、数据加工使用权、数据产品经营权“三权分置”的数据产权制度框架
 - B. 数据所有权统一归国家所有，任何组织和个人不得持有数据
 - C. 数据财产权完全归数据采集者所有，其他主体无权使用
 - D. 个人数据一律禁止进入市场流通，仅可用于公共管理
51. 在数据运营工作中，某数据仓库的 ODS 层（操作数据存储层）每日从业务系统抽取增量数据。若某日抽取任务失败，但源系统数据未发生变化，次日重跑抽取任务时，以下哪种做法最符合数据治理规范？

- A. 直接重跑抽取任务，覆盖前一日未抽取的数据
- B. 先检查源系统数据的时间戳，确认无变化后重跑
- C. 跳过该日数据，从次日开始正常抽取
- D. 删除 ODS 层所有数据，重新全量抽取

二、多选题（共 3 题） 本部分共 3 题。

1. 关于数据挖掘算法在业务场景中的应用，下列描述正确的有：
 - A. 利用 Apriori 算法分析购物篮数据，可以挖掘频繁项集和关联规则
 - B. 使用 K-means 聚类算法进行客户分群，可以帮助识别不同价值特征的用户群体
 - C. 采用决策树算法构建客户流失预测模型，能够输出可解释的规则路径
 - D. 使用朴素贝叶斯算法训练生成对抗网络，用于生成高仿真合成样本
2. 依据 DAMA 数据管理知识体系，下列属于数据治理核心领域的有：
 - A. 数据架构管理
 - B. 数据质量管理
 - C. 元数据管理
 - D. 前端报表可视化开发
3. 根据《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》，下列做法符合我国数据安全与个人信息保护要求的有：
 - A. 对数据实行分类分级保护，对不同级别数据采取对应的安全管理措施
 - B. 向境外提供重要数据，依法通过国家网信部门组织的安全评估
 - C. 处理个人信息具有订立、履行个人作为一方当事人的合同所必需等法定情形的，无需取得个人同意
 - D. 任何情形下，个人均无权请求删除已处理的个人信息

三、判断题（共 5 题） 本部分共 5 题。

1. 数据生命周期管理是指从数据的采集、存储、使用、共享、归档到销毁的全过程进行管控，以确保数据质量、安全与合规。（ ）
2. 根据《中华人民共和国个人信息保护法》，处理敏感个人信息应当取得个人的单独同意，并告知处理敏感个人信息的必要性以及对个人权益的影响。（ ）
3. 依据《数据中心设计规范》（GB 50174-2017），数据中心可根据可用性等级划分为 A、B、C 三级，其中 A 级为容错型，C 级为基本型。（ ）
4. 在“数据资产入表”政策背景下，企业可以将所有数据资源均确认为无形资产或存货，并纳入财务报表。（ ）

5. 在数据生命周期管理中，数据归档后即可不再设置访问控制策略，也无需进行可用性验证。（ ）

四、资料分析（共 2 题） 本部分共 2 题。

1. 国家能源局与国家统计局发布的电力统计数据显示，本组数据覆盖全国规模以上发电企业，装机容量含火电、水电、风电、太阳能及核电全口径，清洁能源指水电、风电、太阳能及核电。用电量按行业分类统计，其中第二产业包括工业与建筑业用电，第三产业包括交通运输、信息技术服务业及商业用电等。在“双碳”目标引导和“十五五”规划筹备背景下，2025 年中央经济工作会议提出加快规划建设新型能源体系，推动能源结构低碳转型。国家发改委、国家能源局联合印发《关于加快建设全国统一电力市场体系的指导意见》，促进清洁能源消纳和电力资源优化配置。2023—2025 年，全国发电装机容量保持较快增长，2025 年达到 38.5 亿千瓦，两年累计增长约 31.8%。清洁能源装机占比从 50.9% 上升至 59.7%，年均提升约 4.4 个百分点。全社会用电量从 92238 亿千瓦时增至 105000 亿千瓦时，年均增速约 6.7%，电力供需总体平衡。分产业看，第二产业用电量仍居首位，但占比从 65.9% 下降至 63.8%，反映制造业结构优化和能耗强度降低；第三产业用电量从 16647 亿千瓦时增至 20300 亿千瓦时，占比提升 1.3 个百分点，成为用电增长的主要驱动力。清洁能源装机占比的提升主要得益于风电、光伏大规模并网，以及特高压输电通道投运。表 1 列出了 2023—2025 年全国发电装机容量、清洁能源装机占比、全社会用电量及分产业用电量的年度数值，便于进行增长率和比重分析（详见下表）。

年份	发电装机容量 (亿千瓦)	清洁能源装机 占比 (%)	全社会用电量 (亿千瓦时)	第二产业用电 量 (亿千瓦时)	第三产业用电 量 (亿千瓦时)
2023	29.2	50.9	92238	60752	16647
2024	33.8	55.3	98500	64010	18230
2025	38.5	59.7	105000	67000	20300

- (1) 2025 年全社会用电量比 2023 年增长约多少？
- A. 12.5%
B. 13.8%
C. 14.2%
D. 15.1%
- (2) 2024 年清洁能源装机容量比 2023 年增加约多少亿千瓦？
- A. 2.94
B. 4.15
C. 3.83
D. 4.60
- (3) 2025 年第三产业用电量占全社会用电量的比重，较 2024 年约提高了多少个百分点？
- A. 0.82
B. 0.68
C. 0.54
D. 0.39
- (4) 2023 年至 2025 年，发电装机容量的年均增长率约为多少？

- A. 12.5%
- B. 14.8%
- C. 13.2%
- D. 16.3%

(5) 若 2026 年清洁能源装机占比比 2025 年再提高 3 个百分点，且总装机容量比 2025 年增长 10%，则 2026 年清洁能源装机容量将比 2025 年增长约多少？

- A. 12.9%
- B. 14.3%
- C. 17.1%
- D. 15.5%

2. 国家能源局数据显示，2022—2025 年，全国电力行业加快绿色转型，全社会用电量持续增长。数据来源于国家能源局定期发布的《全国电力工业统计月报》，统计范围为全国规模以上电力企业，涉及全社会用电量、分产业用电量等指标，数据更新至 2025 年底。在‘碳达峰、碳中和’目标引领下，国家大力推进电能替代，工业、交通、建筑等领域电气化水平不断提升，同时加快构建新型电力系统，风电、光伏等新能源发电装机容量快速增长，为电力消费增长提供了坚实保障。2025 年是‘双碳’目标提出五周年，电力行业清洁低碳转型成效显著，但煤电仍承担兜底保供角色，电力供需呈现‘总体平衡、局部紧张’的态势。

从总体趋势来看，2022—2025 年全社会用电量从 86000 亿千瓦时增至 106000 亿千瓦时，年均增速约 7.2%，各年增速保持在 6.9%—7.6% 的区间，呈先升后降的走势。其中，2024 年全社会用电量同比增速达到约 7.6% 的阶段高点，主要受经济复苏和夏季高温拉动；2025 年增速回落至约 7.1%，但仍保持较快增长。分产业看，第二产业用电量始终占据最大份额，但占比由 2022 年的 66.3% 下降至 2025 年的 64.2%，反映工业结构优化和节能降耗成效；第三产业用电量增速领先，占比由 17.4% 提升至 21.2%，成为用电增长的主要引擎，数字经济和现代服务业的快速发展是重要推动力；第一产业用电量在乡村振兴和农业现代化推动下稳步增长，2025 年达到 1300 亿千瓦时；城乡居民生活用电量 2025 年突破 14000 亿千瓦时，较 2022 年增长约 10.1%，反映出居民生活电气化水平的持续提高。此外，新能源汽车普及、数据中心建设等新业态也增加了用电需求。从绝对量看，第二产业用电量由 2022 年的 57000 亿千瓦时增至 2025 年的 68000 亿千瓦时，增长 19.3%；第三产业用电量由 15000 亿千瓦时增至 22500 亿千瓦时，增长 50.0%；第一产业用电量由 1100 亿千瓦时增至 1300 亿千瓦时，增长 18.2%；城乡居民生活用电量由 12900 亿千瓦时增至 14200 亿千瓦时，增长 10.1%。总体来看，电力消费结构持续优化，绿色低碳转型稳步推进。具体数据见表。

年份	全社会用电量 (亿千瓦时)	第一产业用电 量 (亿千瓦时)	第二产业用电 量 (亿千瓦时)	第三产业用电 量 (亿千瓦时)	城乡居民生活 用电量 (亿千瓦 时)
2022	86000	1100	57000	15000	12900
2023	92000	1150	61000	17000	12850
2024	99000	1200	65000	19500	13300
2025	106000	1300	68000	22500	14200

(1) 2024 年全社会用电量较 2022 年增长了约多少？

- A. 13.5%
- B. 14.6%

C. 16.3%

D. 15.1%

(2) 2023 年第三产业用电量同比增速为多少？

A. 13.3%

B. 11.8%

C. 15.4%

D. 17.6%

(3) 2022—2025 年，第二产业用电量占全社会用电量的比重变化趋势为？

A. 持续上升

B. 先升后降

C. 持续下降

D. 先降后升

(4) 2025 年城乡居民生活用电量较 2022 年增长约多少？

A. 8.5%

B. 9.2%

C. 11.3%

D. 10.1%

(5) 以下说法正确的是（ ）。

A. 2024 年第三产业用电量占全社会用电量的比重比 2023 年提高了约 1.2 个百分点

B. 2022—2025 年，第二产业用电量占全社会用电量的比重持续下降

C. 2023 年城乡居民生活用电量同比下降，降幅约为 0.4%，2024 年同比上升，但 2024 年城乡居民生活用电量仍未超过 2022 年水平

D. 2025 年全社会用电量约是 2022 年的 1.23 倍，年均增速约为 7.6%

参考答案与解析

一、单选题（共 51 题）

1. 答案：A。解析：设原计划需要 x 天完成，则数据总量为 $1200x$ 条。实际每天处理 $1200 \times (1 + 20\%) = 1440$ 条，实际用时为 $(x-2)$ 天。由题意得 $1200x = 1440(x-2)$ ，解得 $x = 12$ ，所以数据总量为 $1200 \times 12 = 14400$ 条。

2. 答案：C。解析：数据资产不同于实物资产，具有非竞争性（多人可同时使用）和部分排他性（可通过技术手段限制访问），其价值在于持续的开发利用，而非消耗。A 项错误，数据可共享；B 项错误，数据使用越多可能价值越大；D 项错误，数据资产的边际效用不一定递减。

3. 答案：A。解析：定义关键词为‘非结构化数据’、‘转化为可量化、可价值评估的结构化数据’。A 项将操作行为（非结构化）转化为风险评分（结构化且有价值评估），符合；B 项只是电子化存储，未结构化；C 项公布已有数据，未转化；D 项统计人数是基础数据汇总，不是从非结构化转化。故答案为 A。

4. 答案：A。解析：每小时增长 15%，即变为原来的 1.15 倍。3 小时后数据量为 $100 \times (1.15)^3 = 100 \times 1.520875 \approx 152GB$ 。

5. 答案：A。解析：数据质量的核心维度通常包括完整性（数据是否缺失）、准确性（数据是否正确）、一致性（数据是否矛盾）和及时性（数据是否及时更新），这是数据治理领域的通用标准。其他选项分别描述了大数据的特征（多样性、海量性等）或数据处理的技术要求，不属于数据质量的直接评价维度。

6. 答案：C。解析：题干中‘清洗’是对‘数据’进行加工处理，去除杂质，为动宾关系，且动作具有加工使其更规范的含义。C 项‘烹饪’是对‘食材’的加工，逻辑最为接近；A 项‘归档’是整理文件，B 项‘完成’是结束任务，D 项‘阅读’是获取知识，均不具有本质加工性质。故答案为 C。

7. 答案：D。解析：总人数 10 人，会 Python 的 6 人，不会 Python 的 4 人。选取 2 人至少一人会 Python 的对立事件是两人都不会 Python，即从 4 人中选 2 人，有 $C(4,2) = 6$ 种；总选法 $C(10,2) = 45$ 种。所以概率 $= 1 - 6/45 = 1 - 2/15 = 13/15$ 。

8. 答案：A。解析：HDFS（Hadoop 分布式文件系统）提供分布式存储，MapReduce 提供分布式批处理计算，是处理海量日志数据的经典组合。MySQL 和 Redis 适用于关系型数据和缓存，MongoDB 和 Elasticsearch 适用于文档存储和搜索，SQLite 和 Excel 适用于小规模数据，均不满足海量分布式处理需求。

9. 答案：B。解析：题干根据‘正相关’得出‘算法优化投入能提升营业收入’的因果结论。B 项指出存在因果倒置，即营收增长是先导，增加投入是结果，直接削弱了题干因果关系；A 项质疑样本代表性，力度较弱；C 项说明部分项目失败，不能否定整体；D 项说明投入比例低，与效果无直接关系。故答案为 B。

10. 答案：B。解析：观察数列， $2 \times 2 - 1 = 3$ ， $3 \times 2 - 1 = 5$ ， $5 \times 2 - 1 = 9$ ， $9 \times 2 - 1 = 17$ ，所以下一项为 $17 \times 2 - 1 = 33$ 。

11. 答案：B。解析：《数据安全法》规定数据处理应当坚持安全与发展并重，采取必要措施保障数据安全，建立健全数据安全治理体系。选项 B 强调开发利用优先于安全保护，与法律精神相悖，因此不包括。

12. 答案：B。解析：定义强调在数据提供方与需求方之间提供服务。B 项是平台利用用户数据进行精准推荐，属于利用数据开展经营活动，并非在供需双方之间撮合并收取服务费，故不属于数据经纪行为。A、C、D 项均在供需间提供撮合、评估、审查或协商服务。故答案为 B。

13. 答案：A。解析：设降价 x 元，售价为 $50-x$ ，销量为 $2000+100x$ 。收入 $= (50-x)(2000+100x) = 100000 + 5000x - 2000x - 100x^2 = 100000 + 3000x - 100x^2$ 。这是开口向下的二次函数，对称轴为 $x = 15$ ，所以售价为 $50 - 15 = 35$ 元时收入最大。

14. 答案: A。解析: UNION 合并两个查询结果并去除重复行; UNION ALL 保留所有行不去重; JOIN 用于连接表; INTERSECT 返回两个查询的交集。因此选 A。

15. 答案: B。解析: 翻译题干: 质量检测 → 清洗; 清洗 → 格式转换, 因此质量检测 → 格式转换, 即所有通过质量检测的数据都通过格式转换, B 项正确。A 项颠倒条件关系, 错误; C 项与题干矛盾; D 项否定前件, 不能推出否定后件, 错误。故答案为 B。

16. 答案: B。解析: 元数据是“关于数据的数据”, 用于描述数据的定义、来源、血缘、质量、使用约束等属性; 主数据是企业核心业务实体数据, 参考数据是标准编码数据, 衍生数据由原始数据加工而来。

17. 答案: B。解析: 前 2 天完成 $15\% \times 2 = 30\%$, 第 3 天完成 20%, 共完成 50%, 剩余 50%。后 2 天平均每天完成 $50\% \div 2 = 25\%$ 。

18. 答案: B。解析: 处理异常值常用方法包括删除、替换、转换等。直接删除可能丢失信息, 忽略或转空值可能影响分析结果, 用均值或中位数替换是常见且稳妥的处理方式, 能减少异常值对整体统计的影响。

19. 答案: A。解析: 谨慎与冒进是反义关系。A 项合规与违规是反义关系, 且二者是中性词与贬义词的对立, 与题干一致; B 项创新与改革是近义关系; C 项效率与质量是并列概念; D 项安全与风险是相反状态, 但并非完全反义。故答案为 A。

20. 答案: C。解析: HDFS 不适合大量小文件, NameNode 压力大; MapReduce 基于磁盘迭代, 适合批处理; YARN 负责集群资源管理和任务调度; Hive 是数据仓库工具, 将 SQL 转换为 MapReduce/Spark 等作业执行, 不是内存计算引擎。

21. 答案: B。解析: 降价前利润率 $= (120 - 80) \div 80 \times 100\% = 50\%$; 降价后售价 $= 120 - 5 = 115$ 元, 利润率 $= (115 - 80) \div 80 \times 100\% = 35 \div 80 \times 100\% = 43.75\%$ 。下降幅度 $= 50\% - 43.75\% = 6.25$ 个百分点。故选 B。

22. 答案: B。解析: 数据销毁是数据生命周期的最后阶段, 主要目的是确保敏感数据被彻底删除, 防止数据泄露, 符合安全合规要求。释放空间和降低成本是附带效应, 但不是核心目的。

23. 答案: A。解析: 乙和丁的陈述构成矛盾关系 (丙使用了 vs 丙没有使用), 必有一真一假。因为只有一人说真话, 则甲和丙说的都是假话。甲说‘我没使用’为假, 说明甲使用了权限; 丙说‘丁使用了’为假, 说明丁没有使用。结合乙丁矛盾, 若乙真则丙使用, 若丁真则丙没用, 两种情况均不冲突, 但甲使用是确定的。因此 A 正确。

24. 答案: A。解析: Z-score 标准化公式为 $(x - \bar{x}) / s$, 处理后的数据均值为 0、标准差为 1。B 是 min-max 归一化, C 是对数变换, D 是标签编码或独热编码。

25. 答案: A。解析: 线性增长, 每天增加 $(30 - 10) / (5 - 1) = 5TB$, 所以第 10 天处理 $10 + (10 - 1) * 5 = 55TB$ 。

26. 答案: A。解析: Apache Kafka 是分布式消息队列, 专为实时数据流处理设计, 支持高吞吐、低延迟。Sqoop 和 DataX 主要用于批量数据迁移, Flume 用于日志采集但实时性不如 Kafka。

27. 答案: C。解析: 建议的根据是培训与绩效正相关。C 项指出培训能够提升数据工具使用熟练度, 为绩效提升提供了内在机制, 直接支持‘增加培训投入’的结论; A 项说明培训内容相关, 但未直接说明对绩效的作用; B 项说明未培训者有经验, 反而削弱; D 项说明高绩效者主动培训, 可能存在反向因果, 削弱。故 C 项最强。

28. 答案: A。解析: 置信度表示在包含 A 的事务中同时包含 B 的条件概率, 即 $\text{confidence}(A \rightarrow B) = \text{support}(A \cup B) / \text{support}(A)$ 。支持度 $\text{support}(A \cup B) / N$ 是支持度, $\text{support}(A) / \text{support}(B)$ 不是置信度。

29. 答案: A。解析: 设月增长率为 r , 则 $1000 * (1 + r)^3 = 2500$, 解得 $(1 + r)^3 = 2.5$, 再过 3 个月为 $2500 * (1 + r)^3 = 2500 * 2.5 = 6250$ 。

30. 答案: C。解析: 无监督学习指训练数据没有标签, K-means 聚类是典型的无监督算法, 用于将数据分成 K 个簇。线性回归、决策树和支持向量机均属于监督学习, 需要标签数据。

31. 答案: B。解析: 由 (2) 成都 = 数据安全; 由 (3) 甲 = 重庆且非数据开发; 又丙 = 数据治理, 则甲不可能是数据治理、数据开发, 也不能是数据安全 (若是安全则甲 = 成都, 与重庆矛盾), 因此甲负责数据运营。工作剩余: 数据安全、数据开发给乙、丁。丁不负责数据开发, 因此丁负责数据安全, 乙负责数据开发。又成都 = 数据安全, 所以丁来自成都。乙不来自贵阳, 且乙负责开发, 剩下城市重庆 (已甲)、成都 (已丁)、贵阳、昆明, 乙不是贵阳, 所以乙来自昆明, 丙来自贵阳。故 B 项正确, A、C、D 均错误。

32. 答案: B。解析: 《个人信息保护法》第二十九条规定, 处理敏感个人信息应当取得个人的单独同意。法律、行政法规规定处理敏感个人信息应当取得书面同意的, 从其规定。

33. 答案: A。解析: 净现值 $= -100 + 30/(1+10\%) + 30/(1+10\%)^2 + \dots + 30/(1+10\%)^5 = -100 + 30 * [1 - (1.1)^{-5}]/0.1 = -100 + 30 * 3.7908 = -100 + 113.72 = 13.72$ 万元。

34. 答案: D。解析: 《个人信息保护法》规定处理个人信息应遵循合法、正当、必要原则, 公开透明原则, 目的限制原则等。选项 D 强调信息自由流通优先, 不符合法律要求。

35. 答案: C。解析: 2025 年比重为 12.5%, 较 2020 年提高 3.2 个百分点, 因此 2020 年比重为 $12.5\% - 3.2\% = 9.3\%$, 故 C 项正确, A 项错误; 比重提高不能推出 GDP 下降, B 项错误; 数字经济核心产业增速与 GDP 增速无直接关系, D 项错误。故答案为 C。

36. 答案: A。解析: 按照规定, 企业日常活动中持有、最终目的用于出售的数据资源, 符合存货定义和确认条件的, 确认为存货; 用于内部研发、管理和运营的数据资源通常确认为无形资产。D 项因无法单独辨认, 不满足资产确认条件。

37. 答案: A。解析: 总选法 $C(6,3) = 20$, 不选 A 和 B 的选法 $C(4,3) = 4$, 所以至少选一个的选法 $20 - 4 = 16$ 。

38. 答案: C。解析: 用户留存率指在一定时间内继续使用产品的用户比例, 直接反映用户留存情况。DAU 和 MAU 反映活跃规模, 用户获取成本反映市场投入, 均不直接衡量留存。

39. 答案: A。解析: 题干矛盾: 流量激增但 CPU 平均利用率下降。A 项说明节点数量大幅增加, 总负载虽增但平均到每个节点的负载下降, 直接解释; B 项暂停非核心业务只能减少部分负载, 但流量激增仍需大量计算, 不足以解释; C 项说明平时被离线任务拉高, 但未解释双十一为何下降, 且流量激增时通常任务更多; D 项单次请求消耗降低, 但总请求量激增, 总消耗未必下降, 解释力度弱。故答案为 A。

40. 答案: A。解析: Spark 利用 RDD 和 DAG 调度、内存计算, 迭代计算性能通常优于 MapReduce。Spark 支持批处理、流处理、交互式查询等场景; MapReduce 是简单的 Map/Reduce 两阶段模型, Spark 直接用 DAG 调度; Spark 可直接读写 HDFS。

41. 答案: B。解析: 设原计划需要 x 天完成, 则数据总量为 $1200x$ 条。实际每天处理 $1200 \times (1+20\%) = 1440$ 条, 实际用时为 $(x-2)$ 天。根据总量不变, 可得 $1200x = 1440(x-2)$, 解得 $x = 12$ 天, 所以数据总量为 $1200 \times 12 = 14400$ 条。

42. 答案: C。解析: Parquet 是列式存储格式, 具有高压缩比和高效的列查询性能, 适合数据仓库场景。文本文件和 CSV 是行式存储, 压缩比低, 查询效率差; 序列文件是 Hadoop 的二进制格式, 但列式查询性能不如 Parquet。

43. 答案: B。解析: 论点是挂钩能提高数据质量。如果员工没有能力或权限改善数据质量, 那么即使挂钩也无法提高, 因此必须假设员工有能力和权限去改善, B 项是必要前提。A 项指指标不透明, 不可行; C 项强调技术依赖, 削弱员工作用; D 项指出可能虚报, 均不能作为前提。故答案为 B。

44. 答案: C。解析: A 更贴近有效性或规范性, B 更贴近准确性, C 正确描述了一致性, D 中“越

短越好”过于绝对，及时性应满足业务时效要求。

45. 答案：C。解析：第 1 天 DAU 为 1000 人，第 2 天为 $1000 \times 1.1 = 1100$ 人，第 3 天为 $1100 \times 1.1 = 1210$ 人，第 4 天为 $1210 \times 1.1 = 1331$ 人，第 5 天下降 5%，则为 $1331 \times 0.95 = 1264.45$ 人。

46. 答案：B。解析：折线图适合展示数据随时间变化的趋势，能清晰显示增减和波动。饼图用于占比，散点图用于相关性，雷达图用于多维度比较，均不适合时间序列趋势展示。

47. 答案：B。解析：题干从‘促销活动后销售额上升’这一时间先后顺序，直接推出促销是原因，典型的‘事后归因’谬误，因此缺陷是错把先后关系当因果关系，B 项正确。A 项说没有相关性不符合题干‘每逢推出促销当月上升’的表述；C 项不是根本缺陷；D 项也是可能缺陷，但相对于 B 项，D 项是‘存在他因’，而题干主张强加因果，B 项更准确地地点明逻辑谬误。故答案为 B。

48. 答案：A。解析：数据生命周期通常从采集开始，经过存储、处理、分析、共享等环节，最后归档并在到期后销毁。C 和 D 的阶段顺序不符合常规逻辑，B 将处理置于存储之前不够合理。

49. 答案：C。解析：《数据安全法》规定，核心数据是指关系国家安全、国民经济命脉、重要民生、重大公共利益的数据。一般工业数据、公共数据和个人一般信息不属于核心数据。

50. 答案：A。解析：“数据二十条”创新性提出建立数据资源持有权、数据加工使用权、数据产品经营权“三权分置”的数据产权制度框架。数据并非一律归国家或采集者独占，个人数据在合法合规前提下也可以依法流通。

51. 答案：B。解析：数据治理规范要求确保数据完整性、准确性和可追溯性。当日抽取失败，次日重跑时，应先检查源系统数据的时间戳或变更日志，确认无变化后重跑，避免重复或遗漏数据。直接重跑可能造成数据重复；跳过会导致数据缺失；全量抽取成本高且可能影响性能。因此 B 选项最符合规范。

二、多选题（共 3 题）

1. 答案：ABC。解析：Apriori 是关联规则经典算法；K-means 是常见聚类算法；决策树模型具有可解释性，能输出决策路径。朴素贝叶斯是概率分类器，不用于训练生成对抗网络。

2. 答案：ABC。解析：DAMA-DMBOK 中数据管理核心领域包括数据治理、数据架构、数据质量管理、元数据管理、数据安全等。前端报表可视化开发属于应用开发，不属于数据治理核心领域。

3. 答案：ABC。解析：《数据安全法》要求对数据分类分级并采取相应保护措施；向境外提供重要数据依法应当进行安全评估；《个人信息保护法》除同意外还规定了订立履行合同所必需等合法处理情形。个人撤回同意等情况下有权请求删除个人信息，故 D 错误。

三、判断题（共 5 题）

1. 答案：正确。解析：数据生命周期管理覆盖数据从产生到消亡的全过程，各阶段均有相应的管理要求，包括采集、存储、使用、共享、归档、销毁等环节，旨在保障数据质量、安全与合规，题干表述正确。

2. 答案：正确。解析：《个人信息保护法》第二十九条要求处理敏感个人信息应当取得个人的单独同意，第三十条要求告知处理敏感个人信息的必要性以及对个人权益的影响，题干表述正确。

3. 答案：正确。解析：GB 50174-2017 将数据中心划分为 A、B、C 三个等级，A 级为容错型，B 级为冗余型，C 级为基本型，题干表述正确。

4. 答案：错误。解析：并非所有数据资源都能入表，只有符合资产定义和确认条件的数据资源才能确认为无形资产或存货，需同时满足企业拥有或控制、预期带来经济利益等条件。题干说法过于绝对，因此错误。

5. 答案：错误。解析：数据归档后仍属于组织数据资产的一部分，必须继续设置访问控制策略，并定期进行可用性验证，以满足合规审计和数据安全要求。

四、资料分析（共 2 题）

1. 答案: B; C; A; B; D

(1) B. $(105000 - 92238) \div 92238 = 12762 \div 92238 \approx 13.84\%$, 最接近 13.8%。

(2) C. 2023 年清洁能源装机容量 = $29.2 \times 50.9\% \approx 14.86$ 亿千瓦; 2024 年 = $33.8 \times 55.3\% \approx 18.69$ 亿千瓦; 增加 $18.69 - 14.86 = 3.83$ 亿千瓦。

(3) A. 2025 年比重 = $20300 \div 105000 \approx 19.33\%$; 2024 年比重 = $18230 \div 98500 \approx 18.51\%$; 提高 $19.33 - 18.51 = 0.82$ 个百分点。

(4) B. 设年均增长率为 r , 则 $(1+r)^2 = 38.5 \div 29.2 \approx 1.3185$, 开平方得 $1+r \approx 1.1482$, $r \approx 14.82\%$, 约 14.8%。

(5) D. 2025 年清洁能源装机容量 = $38.5 \times 59.7\% \approx 22.98$ 亿千瓦; 2026 年总装机 = $38.5 \times 1.1 = 42.35$ 亿千瓦; 2026 年清洁能源装机占比 = $59.7\% + 3\% = 62.7\%$; 2026 年清洁能源装机容量 = $42.35 \times 62.7\% \approx 26.55$ 亿千瓦; 增长率 = $(26.55 - 22.98) \div 22.98 \approx 15.53\%$, 约 15.5%。

2. 答案: D; A; B; D; A

(1) D. $(99000 - 86000) \div 86000 = 13000 \div 86000 \approx 0.1512$, 即 15.1%。故选 D。

(2) A. $(17000 - 15000) \div 15000 = 2000 \div 15000 \approx 0.1333$, 即 13.3%。故选 A。

(3) B. 2022 年比重 = $57000 \div 86000 \approx 66.28\%$, 2023 年比重 = $61000 \div 92000 \approx 66.30\%$, 2024 年比重 = $65000 \div 99000 \approx 65.66\%$, 2025 年比重 = $68000 \div 106000 \approx 64.15\%$, 可见先升后降。故选 B。

(4) D. $(14200 - 12900) \div 12900 = 1300 \div 12900 \approx 0.1008$, 即 10.1%。故选 D。

(5) A. A 项, 2022 年第二产业占比 = $57000 \div 86000 \approx 66.28\%$, 2023 年占比 = $61000 \div 92000 \approx 66.30\%$, 2024 年占比 = $65000 \div 99000 \approx 65.66\%$, 2025 年占比 = $68000 \div 106000 \approx 64.15\%$, 可见占比先升后降, 并非持续下降, A 错误。B 项, 2023 年城乡居民生活用电量同比下降 $(12900 - 12850) \div 12900 \approx 0.39\%$, 2024 年同比上升 $(13300 - 12850) \div 12850 \approx 3.50\%$, 且 2024 年 13300 亿千瓦时已超过 2022 年的 12900 亿千瓦时, B 中“仍未超过 2022 年水平”说法错误, 故 B 错误。C 项, 2024 年第三产业用电量占比 = $19500 \div 99000 \approx 19.70\%$, 2023 年占比 = $17000 \div 92000 \approx 18.48\%$, 提高了 $19.70\% - 18.48\% = 1.22$ 个百分点, 约 1.2 个百分点, C 正确。D 项, 2025 年全社会用电量是 2022 年的 $106000 \div 86000 \approx 1.23$ 倍, 年均增速 = $(106000 \div 86000)^{\frac{1}{3}} - 1 \approx 7.2\%$, 不是 7.6%, D 错误。故选 A。