## 大数据基础复习资料答案

### 一、客观题答案

1. B 2. D 3. B 4. C 5. B 6. D 7. C 8. A 9. B 10. A
2. B 12. A 13. B 14. A 15. A 16. C 17. B 18. B 19. C 20. D

### 二、判断题答案

1. × 2. √ 3. × 4. × 5. √ 6. × 7. √ 8. × 9. √ 10. ×
2. √ 12. × 13. √ 14. √ 15. × 16. √ 17. × 18. √ 19. × 20. √

### 三、简答题答案

1. 数据的定义：在拉丁文里意为 “已知”“事实”，《牛津英语字典》定义为计算机执行操作的量、字符或符号，以电信号存储传输并记录在特定介质上；《新华字典》定义为数字、文字、字母等计算机加工处理的对象，可统指计算机中所有电子化记录的信息。常见类型划分角度：原始数据是否为直接来源、收集过程是否有控制因素、是否与时间相关、结构化程度等。
2. 大数据的处理流程主要包括数据的采集与预处理、存储与管理、可视化、分析与挖掘等步骤，可根据实际需求进行迭代。
3. 大数据在政府领域的主要应用：①政府管理服务：收集治理过程中的结构化和非结构化数据，对相关要素进行全过程精确分析，为政府决策提供数据支撑，为民众提供精准服务；②政务信息服务共享：打造大数据施政平台，整合分散在不同部门、行业的数据，打破共享壁垒，优化服务流程，促进协同治理，提高决策及时性。
4. 数据预处理的目的：解决原始数据格式不一致、散乱、缺失值、异常值和高维度等问题，提高数据质量，使数据形态符合后期统计模型或数据挖掘算法的要求，以便更好地挖掘数据价值。主要操作包括数据整理（提取、聚合、连接、去除冗余和重复）、数据变换（标准化、对数变换等）、缺失值处理、异常值检测、数据降维。
5. HDFS 的主要特点：①数据复制：确保数据始终可用，防止丢失，故障时可从其他节点获取备份；②容错性和可靠性：集群由多台服务器组成，跨节点存储数据，保障系统稳定；③高可用性：单个节点故障不影响数据访问；④可扩展性：可根据需求轻松扩展集群节点；⑤高吞吐量：分布式存储实现并行数据处理，减少处理时间。
6. 数据可视化的基本原则：①“信”：可视化效果准确表达数据信息，不产生偏差或歧义；②“达”：通过图表清晰传递数据信息；③“雅”：设计出令人赏心悦目的可视化图表。
7. 分类问题的定义：从数据集中学习一个目标函数，应用于检验集，根据数据对象的特征判断其归属的目标类别，训练集类别标号已知，用于建立模型，检验集类别标号未知，用于评价模型性能。分类算法主要类型：基于示例的分类算法（如 k 近邻算法）、基于概率模型的分类算法（如朴素贝叶斯分类器）、基于决策树模型的分类算法（如 C5.0、随机森林）、基于线性模型的分类算法（如支持向量机）等。
8. 大数据隐私保护的对策主要来自法律、技术和管理三个方面：①法律措施：国家制定和完善数据隐私保护相关法律法规；②技术措施：采用各类隐私保护技术；③管理措施：国家、企业和个人三个层面分别采取相应的管理手段，加强隐私保护。

### 四、综合题答案

1. 大数据技术在电子商务领域的应用场景及提升竞争力的方式：

应用场景：①精准营销：根据用户的搜索关键词、购买记录、浏览行为等数据，构建用户画像，推送个性化商品推荐；②客户关系管理：通过分析客户的购买频率、消费金额、评价内容等，进行客户生命周期管理，针对不同阶段客户制定差异化策略；③库存管理：基于历史销售数据、市场趋势数据等，预测商品需求，优化库存配置，减少库存积压和缺货情况；④舆情监测：监测社交网络、商品评价等渠道的舆情信息，及时了解消费者对商品和服务的反馈，处理负面舆情，改进产品和服务。

提升竞争力的方式：①提高用户体验：个性化推荐让用户快速找到心仪商品，提升购物便捷性和满意度，增加用户粘性；②降低运营成本：精准的需求预测优化库存，减少资金占用和仓储成本，舆情监测及时发现问题，降低品牌损失；③提升营销效果：精准营销提高广告投放转化率，降低营销成本，同时根据客户反馈快速调整营销策略，抢占市场先机；④增强品牌口碑：通过优化产品和服务、及时处理客户诉求，树立良好的品牌形象，吸引更多潜在客户。

1. MapReduce 的编程模型与执行流程及应用实例：

编程模型：以一组键值对（key，value）作为输入，经过 Mapping 和 Reducing 两个阶段的计算，输出另一组键值对。开发者需定义 Map 函数和 Reduce 函数，底层系统自动完成资源部署和并行计算。

执行流程：①Mapping 阶段：Mapper 类将不同数据节点中的数据进行映射，把输入的键值对转换为中间键值对；②Reducing 阶段：Reducer 类接收 Mapper 输出的中间键值对，进行整合、汇总，生成最终的键值对输出。

应用实例：统计文本内容中每个单词出现的次数。①Mapping 阶段：将文本数据按行分割，每个单词作为 key，value 设为 1，例如 “Dear Bear River” 映射为（Dear,1）、（Bear,1）、（River,1）；②Reducing 阶段：将相同 key 的 value 进行求和，例如将所有（Bear,1）汇总为（Bear,2），最终得到每个单词的总出现次数。该应用可用于海量文本数据的词频统计，如搜索引擎的关键词分析、文本内容主题挖掘等场景。