**特许全球金融科技师（CGFT）一级**

**《大数据技术原理及应用》考试大纲**

（修订日期： 2022年09月21日）

[一、 考试性质 2](#_Toc114444094)

[二、 考查目标 2](#_Toc114444095)

[三、 考试形式和试卷结构 3](#_Toc114444096)

[1. 试卷满分 3](#_Toc114444097)

[2. 答题方式 3](#_Toc114444098)

[3. 试卷内容结构 3](#_Toc114444099)

[4. 试卷题型结构 3](#_Toc114444100)

[四、 考察内容与考点解析 4](#_Toc114444101)

[1. 章节内容考点清单 4](#_Toc114444102)

[2. 重要高频考点提示 14](#_Toc114444103)

[五、 历年真题与考试样题 15](#_Toc114444104)

# 考试性质

特许全球金融科技师（CGFT）一级包含《大数据技术原理及应用》科目，大数据技术原理及应用是金融科技从业基础知识和技能，特许全球金融科技师（CGFT）一级考试是技能认证性考试，按照金融科技人才知识技能能力标准，以标准化考试方式对考生对于大数据知识技能的掌握程度进行评价。

# 考查目标

1. **知识技能要求**

CGFT一级对考生大数据技术原理及应用相关知识技能要求如下：

* 理解数据库和数据仓库基本概念
* 理解关系数据库原理
* 熟练使用 SQL 进行数据查询和数据分析
* 深入了解数据挖掘基本原理、算法模型和商业应用
* 深入了解数据挖掘基本原理、算法模型和商业应用
* 了解自然语言处理基本原理、算法模型和商业应用
* 了解知识图谱基本原理、算法模型和商业应用
* 了解 MySQL 和 Oracle 等数据库产品应用
* 深入了解金融大数据场景应用和实践案例
* 了解 Bloomberg 和 Wind 金融数据库终端产品使用

1. **能力素质要求**

CGFT一级对考生大数据技术原理及应用相关能力素质要求如下：

* 能深刻理解关系数据模型，并进行数据模型设计
* 能灵活应用 SQL 语言进行数据处理、查询和分析，并解决实际业务问题
* 具备灵活使用数据挖掘、自然语言处理、知识图谱工具解决金融风控、营销、运营等场景问题
* 具备一定的大数据应用创新能力和解决方案设计能力

# 考试形式和试卷结构

## 试卷满分

CGFT一级考试包含七门课程，其中《大数据技术原理及应用》包含40题，该部分满分为100分。

## 答题方式

机考，闭卷，可以跨区答题

## 试卷内容结构

不对内容结构分布做明确限定。

## 试卷题型结构

以客观题方式进行考察，题型包括：判断题、单选题、多选题。

不对题型结构分布做明确限定。

# 考察内容与考点解析

## 章节内容考点清单

本表中考察要求的定义如下

* 基础识记：主要是事实知识，考察对事实性知识的记忆，“知道”即可。
* 一般理解：主要是概念定义，仅需一般理解。
* 深入理解：主要是核心原理，需要较深入理解并能灵活应用。
* 不做要求：不作为考试内容，考生可以自主选择阅读。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 章 | 节 | 考点 | 考查要求 | 重要性 | 难度 |
| 1. 数据库基本理论 | 1.1 数据库管理系统DBMS | * 数据库管理系统（DBMS）是什么 * 文件系统储存数据的弊端 * 三层抽象：物理层、逻辑层、视图层 * 实例与模式，模式分为物理模式和逻辑 模式 * 数据模型的定义和分类： 关系模型、实体-联系数据模型、基于对象的数据模型、半结构化的数据模型 | 基础识记 | 高 | 低 |
| 1.2 关系数据库 | * 数据定义语言 DDL 基本概念，数据字典 的基本概念 * 数据操纵语言（DML） * 非过程化的：用户指定需要什么数据而不必指定如何获取这部分数据（比如 SQL） * 过程化的：用户指定需要什么数据以及如何获取这部分数据（比如 relational algebra, SQL procedural extensions） * SQL 是最广泛使用的查询语言 | 基础识记 | 高 | 低 |
| 1.3 数据库设计 | * 数据库设计的过程：概念设计🡪逻辑设 计🡪物理设计 | 基础识记 | 中 | 低 |
| 1.4 存储管理 | * 数据库系统的功能可分为：存储管理器 + 查询处理器 * 存储管理器是数据库系统中负责在数据库中存储的底层数据与应用程序以及向系统提交的查询之间提供接口的部件 * 存储管理器负责的任务：与文件管理器进行交互、有效的存储、检索和更新 * 存储管理难点：存储访问、文件管理、索引和散列 | 基础识记 | 高 | 低 |
| 1.5 事务管理 | * 事务是在数据库应用程序中执行单个逻辑功能的操作的集合 * 事务管理组件可确保即使出现系统故障（例如电源故障和操作系统崩溃）和事务故障，数据库仍保持一致（正确）状态 | 基础识记 | 高 | 低 |
| 1.6 数据库系统结构 | * 数据库结构分为：集中式、C/S 结构、并行分布式 | 基础识记 | 低 | 低 |
| 1.7 数据库系统的历史 | * 各年代数据库技术 | 基础识记 | 低 | 低 |
| 1. 2. 关系数据库 | 2.1 关系模型的介绍 | * 关系数据库的结构 （关系、属性、关系模式、关系属性、码、外码、模式图） * 查询语言分类：过程化语言、非过程化用语言 * 基础关系代数操作，六种基本运算 | 深入理解 | 高 | 高 |
| 2.2 关系数据库查询语言 SQL | * SQL 查询语言的基本概念 * SQL 常用语句，数据定义、创建表、查询、基本运算、集合运算、聚集函数、嵌套、数据库的修改等 | 掌握操作 | 高 | 中 |
| 2.3 关系数据库查询语言 SQL （中级） | * 连接关系、视图、事务、完整性约束 * SQL 数据类型和模式 * 授权 | 掌握操作 | 高 | 高 |
| 1. 3. 数据库设计 | 3.1 数据库设计和实 体关系模型 | * 元组关系演算、域关系演算 * 数据库可以被建模成: 实体, 实体间的关系 * 实体集合、关系集合、属性 * 二元关系 * 实体-联系图（E-R图） | 深入理解 | 高 | 高 |
| * 1. 3.2 关系数据库设计 | * 好的关系设计的特点 * 使用函数依赖的分解 * 使用多值依赖的分解 * 更多范式 * 数据库设计过程 | 一般理解 | 低 | 高 |
| 1. 4. 数据存储 | 4.1存储和文件结构 | 物理存储介质   * 高速缓存存储器 * 主存储器 * 快闪存储器 * 磁盘存储器 * 光盘存储 * 磁带存储 * 磁盘、磁盘（性能度量标准） * RAID * 文件组织 * 数据字典 | 基础识记 | 高 | 中 |
| 4.2索引和散列 | * 搜索码、顺序索引、散列索引、B+树索引 | 基础识记 | 中 | 低 |
| 1. 5. 数据查询 | 5.1事务查询 | * 查询处理基本步骤：解析, 翻译，评估 * 查询代价的度量 * 文件扫描 （线性搜索） * 索引扫描（主索引、等值比较） * 外部归并排序：两个阶段 – 1）创建 2）归并 | 基础识记 | 高 | 低 |
| 5.2查询优化 | * 查询优化 * 代价估计 | 一般理解 | 中 | 低 |
| 1. 6. 事务管理 | 6.1事务与事务管理 | * 事务概念及面临的两个主要问题：故障、并发 * 事务的ACID性质 * 事务状态 * 并发执行 * 调度 * 可串行化 * 可恢复调度 * 级联回滚 | 深入理解 | 高 | 低 |
| 6.2隔离级别 | * 并发控制 * 隔离级别 * 隔离的实现 | 一般理解 | 中 | 低 |
| 1. 7. 数据挖掘 | 7.1 OLAP 与数据仓库 | * 数据库应用包括：事务处理系统、决策支持系统 * 决策支持系统的基本概念和定义 * 联机分析处理 OLAP 的基本概念 * 分析处理基本操作：转轴、切片、上卷、下钻 * 交叉表/透视表基本概念 * 数据仓库的基本概念 * 如何收集数据：源驱动、目的驱动 | 基础识记 | 高 | 低 |
| 7.2数据挖掘 | * 预测性模式：分类、回归 * 描述性模式：关联、聚类 | 一般理解 | 高 | 低 |
| 1. 8. 自然语言处理 | 8.1 词法分析 | * 语言分析层次 * 词是基于最小的语义单元-词素而构成的 * 使用有限状态自动机(Finite State Automata)来表示词的形成过程 * 自动分词就是让计算机系统在文本中的词与词之间自动加上空格或其他边界标记 | 基础识记 | 高 | 低 |
| 8.2 句法分析 | * 解析器的概念 * 成分语法 * 依赖语法 | 基础识记 | 低 | 低 |
| 8.3 语义分析 | * 语义角色标注 * 指代消解 * 情感分析 * 词向量 | 深入理解 | 高 | 高 |
| 8.4 信息检索 | * 信息检索 * WEB 检索 | 基础识记 | 低 | 高 |
| 1. 9. 知识图谱 技术知识 与案例 | 9.1什么是知识图谱 | 基本概念   * + - 知识图谱发展历史     - 知识图谱基本概念     - 为什么使用知识图谱     - 文本中的信息抽取面临挑战 | 一般理解 | 高 | 中 |
| 9.2文本中知识抽取 | * 什么是知识抽取 * 实体、属性、关系 | 一般理解 | 高 | 中 |
| 9.3知识图谱应用 | * 知识推理 * 知识表示 * 常见知识图谱系统 | 一般理解 | 高 | 中 |
| 1. 10. 常用数据库产品应用 | 10.1 MySQL 数据库 | * MySQL 的优势 | 基础识记 | 高 | 中 |
| 10.2 Oracle 数据库 | * Oracle 的使用 * 与 MySQL 区别 | 基础识记 | 高 | 中 |
| 10.3 NoSQL 数据库 | * NoSQL 数据库基本概念及其应用 | 基础识记 | 高 | 中 |
| 1. 11. 金融大数据应用 | 11.1 金融大数据的发展 | * 金融大数据的兴起、产生、获取 * 金融大数据发展特点 * 金融大数据的发展趋势 | 基础识记 | 低 | 低 |
| * 1. 金融大数据的应用前景 | * 金融大数据的应用前景 ：银行、保险、证券、支付、互联网金融 * 云服务 | 基础识记 | 中 | 低 |
| * 1. 金融大数据的典型案例分析 | * 风险控制 * 精准营销 | 基础识记 | 中 | 低 |
| 11.4 挑战与建议 | * 挑战及相关建议 | 一般理解 | 低 | 低 |
| 12. 常用金融数据库实务 | 12.1 Bloomberg | 概述   * 红色、绿色、黄色操作键 * 面板功能 | 基础识记 | 中 | 低 |
| 12.2 WIND | 功能使用  新闻公告、全球行情、数据资料、投资分析 | 基础识记 | 中 | 低 |

## 重要高频考点题目

* 数据库管理系统(DBMS)基本概念
* 数据库系统特点
* 数据的抽象视图
* 数据结构分类
* 数据模型
* 关系数据库的基本数据结构
* SQL查询语言
* 数据库建模
* 数据存储
* 索引
* 数据挖掘基本概念
* 自然语言处理概念
* 知识图谱概念
* MySQL、Oracle、NoSQL 数据库应用
* 金融大数据应用

# 历年真题与考试样题

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 题面 | 选项列 | | | | 正确答案 |
| A | B | C | D |  |
| 选出下列不需要数据库的软件 | 银行管理系统 | 学生成绩管理系统 | 酒店管理系统 | 计算器 | D |
| 数据库（DB），数据库系统（DBS）和数据库管理系统（DBMS）之间的关系是 | DBMS包括DB和DBS | DBS包括DB和DBMS | DB包括DBS和DBMS | DBS就是DB，也就是DBMS | B |
| 提供数据库定义、数据操纵、数据控制和数据库维护功能的软件称为 | OS | DS | DBMS | DBS | C |
| 下列四项中，不属于数据库系统特点的是 | 数据结构化 | 数据由DBMS统一管理和控制 | 数据冗余度高 | 数据独立性高 | C |
| 数据的抽象视图不包括哪些层级 | 物理层 | 逻辑层 | 应用层 | 视图层 | C |
| XML可扩展标记语言是结构化数据模型。 | 是 | 否 |  |  | B |
| 大学教师与课程的联系类型是 | 一对一 | 一对多 | 多对多 | 多对一 | C |
| 出版社与图书的联系类型是 | 一对一 | 一对多 | 多对多 | 多对一 | B |
| 属性值不一定是原子的，也就是可分割的。 | 是 | 否 |  |  | B |
| 关系name×city的子集可能是 | （李华, 张波） | （李华，北京） | （北京，上海） | （上海，张波） | B |
| 关系数据库的基本数据结构是 | 树 | 图 | 索引 | 关系 | D |
| 有一名列车运营实体，含有：车次，日期，实际发车时间，实际抵达时间，情况摘要等属性，该实体主码是 | 车次 | 日期 | 车次和日期 | 车次和情况摘要 | C |
| 哪个不是纯查询语言 | 关系代数 | 元组关系验算 | 域关系验算 | 键码验算 | D |
| 已知关系R和S，则R∪S等价于 | (R-S)-S | S-(S-R) | (R-S)+S | S-(R+S) | C |
| 选择运算的谓词不包括 | and | or | > | ∈ | D |
| 运算σ\_(amount>1200) (loan)的含义是找出： | 贷款额度超过120的数额 | 贷款额度超过1200的帐号 | 贷款额度不超过1200的帐号 | 贷款额度超过1200的数额 | B |
| 已知关系R和S，则R∩S等价于 | (R-S)-S | S-(S-R) | (S-R)-R | S-(R-S) | B |
| 元组关系演算是一种过程查询语言 | 是 | 否 |  |  | B |
| 查询语句{ t | ∃ s ∈ instructor (t [ID] = s [ID] ∧ s [salary] > 8000)}的含义是 | 薪资大于8000 的导师ID，姓名，部门名称 | 薪资大于8000 的导师ID | 薪资小于8000 的导师ID，姓名，部门名称 | 薪资小于8000 的导师ID | B |
| 两组非过程查询语言是指 | 定义域关系验算和元素关系验算 | 定义域关系验算和元组关系验算 | 定点关系验算和元素关系验算 | 定点关系验算和元组关系验算 | B |
| 下列哪一个不是在描述表达式{t | P(t)}的安全性 | 可以生成无限关系的元组演算表达式 | 确保表达式安全性的表达式集 | 防止生成出无限关系 | 元组t的每个分量都出现在P的关系、元组或者常量之中 | A |
| 反映现实世界中实体及实体间联系的信息模型是 | 关系模型 | 层次模型 | 网状模型 | E-R模型 | D |
| 下列存储结构最稳定的是 | 主存储器 | 快闪存储器 | 磁带存储 | 磁盘存储 | C |
| 以下哪个存储介质是易丢失存储 | 主存储器 | 快闪存储器 | 磁带存储 | 磁盘存储 | A |
| 磁盘性能的度量标准不包括 | 访问时间 | 数据传输速率 | 可靠性 | 体积 | D |
| 储存量最大的存储介质是 | 磁带储器 | 快闪存储器 | 光盘存储 | 磁盘存储 | A |
| 磁盘的可靠性主要反映在平均故障时间，也即磁盘平均可以连续运行而不会发生任何故障的平均时间。 | 是 | 否 |  |  | A |
| 对于磁盘来说，所有盘片的第i个磁道合在一起成为以下哪个磁盘结构 | 第i个扇区 | 第i个磁头 | 第i个磁道 | 第i个柱面 | D |
| 一个新磁盘的MTTF为1200000小时，这意味着给定10000个相对较新的磁盘，出现故障的间隔是： | 1200小时 | 120小时 | 2400小时 | 240小时 | B |
| 多个磁盘可能共享一个控制器，以下控制器数据处理速率最快的是 | SATA | SCSI | SAS | ATA | C |
| 如今RAID被用于更高的可靠性和带宽。“I”的含义是： | 经济 | 高效 | 独立 | 可靠 | C |
| 对于查询代价的度量，最主要的代价是 | 磁盘存取 | CPU执行时间 | 网络通信时间 | 寻找等效表达式 | A |
| 计算磁盘访问的代价，不需要考虑的指标是 | 磁盘搜索次数 | 读取的块数 | 网络通信时间 | 写入的块数 | C |
| 在进行选择查询时，关于文件扫描方法的描述不正确的是 | 其本质是线性搜索 | 可以进行二进制搜索 | 需要扫描每个文件块并测试所有记录 | 如果选择位于关键属性上，则可以停止查找记录 | B |
| 使用索引的搜索算法称为索引扫描，其不包括： | 线性搜索 | 主索引等值比较 | 码属性等值比较 | 非码属性等值比较 | A |
| 在数据查询时需要面对复杂选择，下列哪一个无法实现合取选择 | 利用一个索引的合取选择 | 通过标识符的并集进行合取选择 | 使用复合索引进行合取选择 | 通过标识符的交集进行合取选择 | B |
| 下列索引算法中代价最昂贵的是 | 检索满足相应相等条件的单个记录 | 搜索键不是候选键，检索多个记录 | 检索多个记录 | 搜索键是候选键，检索单个记录 | B |
| 外部归并排序有几个阶段 | 1 | 2 | 3 | 4 | B |
| 事务需要面对的问题是 | 计算机出现故障 | 事务本身失败 | 多个事务并行 | 以上都是 | D |
| 事务的ACID特性中I是指 | 事务对数据库的更新必须是永久的 | 事务执行过程中要么全部执行成功要么全部执行失败 | 并发执行的多个事务互不影响 | 事务成功完成后，数据库必须保持一致 | C |
| 事务（Transaction）是一个 | 程序 | 进程 | 操作序列 | 完整性规则 | C |
| 数据库事务的ACID特性，不包括 | 原子性 | 事务性 | 一致性 | 隔离性 | B |
| 事务的ACID特性中C是指 | 事务对数据库的更新必须是永久的 | 事务执行过程中要么全部执行成功要么全部执行失败 | 并发执行的多个事务互不影响 | 事务成功完成后，数据库必须保持一致 | D |
| 关于事务的中止状态，说法不正确的是 | 目的是使数据库恢复到事务开始前的状态 | 需要撤销事务对数据库的改变 | 中止后选择重启事务 | 中止后选择杀死事务 | C |
| 事务的执行次序称为 | 过程 | 步骤 | 调度 | 优先级 | C |
| 哪一项不是并发执行的优势 | 提高处理器和磁盘利用率 | 保持数据库的一致性 | 增加系统吞吐量 | 减少事务平均响应时间 | B |
| 使用一个属性的固定值查看交叉表（单维）的过程为 | 切片 | 切块 | 转轴 | 切片和切块 | A |
| OLAP的操作涉及计算一个或多个维度的所有数据关系 | 上卷 | 下钻 | 切块 | 转轴 | A |
| 可视化表示的交叉表的一般化是\_\_\_，也称为数据立方体。 | 二维立方体 | 多维立方体 | N维立方体 | 长方体 | A |
| 语言分析的层次不包括 | 词汇分析 | 句法分析 | 语义分析 | 结构分析 | D |
| 有限状态机中定义的状态不包括 | 开始 | 结束 | 中间 | 重心 | D |
| 最小语义单元 | 词 | 词元 | 词素 | 词根 | C |
| 让计算机系统在文本中的词与词之间自动加上空格或其他边界标记 | 分词 | 分句 | 词性标注 | 词形还原 | A |
| 基于态度的情感分析不包括 | 态度的持有者 | 态度的目标 | 态度的类型 | 态度的原因 | D |
| 文本中的信息抽取面对的主要挑战不包括 | 分块（chunking） | 分词 | 语义消歧 | 关系抽取 | B |
| 不属于知识图谱中的元数据的是 | 作者的名字 | 书籍的出版社地址 | 书的价格 | 书的内容 | D |
| 以下哪项用来排序 | ORDER BY | ORDERED BY | GROUP BY | GROUPED BY | A |
| 在SELECT、INSERT、UPDATE、DELETE等指令中，最重要也是最频繁使用的语句是 | SELECT | INSERT | UPDATE | DELETE | A |
| 向数据表中插入数据用哪一项 | SELECT | INSERT | UPDATE | DELETE | B |
| 在SELECT语句中，使用何种关键字可以屏蔽重复行 | TOP | ALL | UNION | DISTINCT | D |
| 创建数据库用以下哪一项 | create mytest | create table mytest | database mytest | create database mytest | D |
| SELECT COUNT(SAL) FROM EMP GROUP BY DEPTNO; 语句的意思是 | 求每个部门中的工资 | 求每个部门中工资的大小 | 求每个部门中工资的总数 | 求每个部门中工资的个数 | D |
| 下列哪个数据属于个人行动数据 | 行政处罚数据 | 高管变动数据 | 产品评分数据 | 资产重组数据 | C |
| 混合云架构不包括 | 所有IT系统架构在云端 | 非核心应用迁移到公有云 | 部分核心应用迁移到私有云 | 关键业务继续使用传统架构 | A |
| 关于金融大数据的发展，有前景的发展方向不包括 | 实现个性化创新 | 学科交叉融合 | 数据寡头 | 覆盖长尾用户 | C |
| 下列哪个不属于金融云服务的特点 | 快速交付 | 拓展性强 | 无需租借服务器 | 低运维成本 | C |
| 金融机构最常使用的大数据应用场景中，大数据技术参与度最低的是 | 反欺诈 | 精准营销 | 基金定投 | 交易预警 | C |
| 供应链金融领域主要用到的大数据技术是 | 知识图谱 | 区块链 | 分布式系统 | 数据标注 | A |
| 关于彭博终端说法不妥当的是 | 是一套计算机软件，也是金融资讯的代名词 | 用户可以直接使用终端进行交流 | 为专业人士提供实时金融市场数据和相关服务 | 仅能通过出售的彭博终端访问服务应用 | D |
| 彭博终端的面板不包括哪个部分 | 工具栏 | 键盘 | 命令行 | 功能区域 | B |
| 关于彭博终端键盘说法正确的是 | 绿色表示市场板块键 | 黄色表示操作键 | 红色是停止键 | 键盘与普通计算机键盘相同 | C |
| 万得咨询金融终端的特色功能不包括 | 新闻公告 | 数据分析 | 即时通讯 | 期权交易 | D |
| 在万得咨询金融终端使用F9五部看股，哪一项是其中一部 | 交易数据 | 重要股东 | 财务数据 | 新闻公告 | C |
| 在万得咨询金融终端使用F9五部看股，哪一项不是其中一部 | 交易数据 | 同业比较 | 财务数据 | 盈利预测 | A |

五步看股法： Ø 公司资料 Ø 财务数据 Ø 重大事件 Ø 同业比较 Ø 盈利预测