张家界航空工业职业技术学院
ZHANGJIAJIE INSTITUTE OF AERONAUTICAL ENGINEERING

大数据技术

专业技能考核题库

专 业 名 称:	大数据技术
专 业 代 码:	510205
适 用 年 级:	2021级
所 属 学 院:	信息技术学院
专业负责人:	魏红伟
制(修)订时间:	2022年4月

目录

张家界航空工业职业技术学院
大数据技术专业技能考核题库

一、专业基本技能模块

模块一 程序设计

1.试题编号：1-1 任务实现1

（1）任务描述

任务一：从键盘读入三个不相同的数，把这三个数由小到大输出(20分)。

要求：使用分支结构语句实现。

任务二：使用循环语句打印出如下图案（30分）。

*

***

*****

****

要求：使用循环结构语句实现。

任务三：从键盘输入x，根据以下情形求y的值（30分）：

y=0;（当x<=0时）

y=2x+1;（当0<x<5时）

y=x2-1;（当x>=5时）

要求：使用多分支条件语句实现。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

2.试题编号：1-2 任务实现2

（1）任务描述

任务一：输入一个百分制分数，输出其对应的五级制成绩，包括：优（90分以上，含90）、良（80-90分，含80，不含90）、中（70-80分，含70不含80）、及格（60-70分，含60不含70）、不及格（60分以下，不含60）(20分)

要求：使用多分支条件语句实现。

任务二：输出阶梯形式的9*9乘法口诀表，如下图所示（30分）。

1*1=1
1*2=2 2*2=4
1*3=3 2*3=6 3*3=9
1*4=4 2*4=8 3*4=12 4*4=16
1*5=5 2*5=10 3*5=15 4*5=20 5*5=25
1*6=6 2*6=12 3*6=18 4*6=24 5*6=30
1*7=7 2*7=14 3*7=21 4*7=28 5*7=35 6*7=42 7*7=49
1*8=8 2*8=16 3*8=24 4*8=32 5*8=40 6*8=48 7*8=56 8*8=64
1*9=9 2*9=18 3*9=27 4*9=36 5*9=45 6*9=54 7*9=63 8*9=72 9*9=81

要求：使用循环结构语句实现。

任务三：输入某人的收入，计算个人应缴的税额。如税率表所示（30分）。

税率表

级数	全月应纳税所得额	适用税率%	速算扣除数（元）
1	不超过500元的	5	0
2	超过500元至2000元的部分	10	25
3	超过2000元至5000元的部分	15	125
4	超过5000元至20000元的部分	20	375
5	超过20000元至40000元的部分	25	1375
6	超过40000元至60000元的部分	30	3375
7	超过60000元至80000元的部分	35	6375
8	超过80000元至100000元的部分	40	10375
9	超过100000元的部分	45	15375

要求：使用多分支条件语句实现。

（2）作品提交要求见本模块附录1

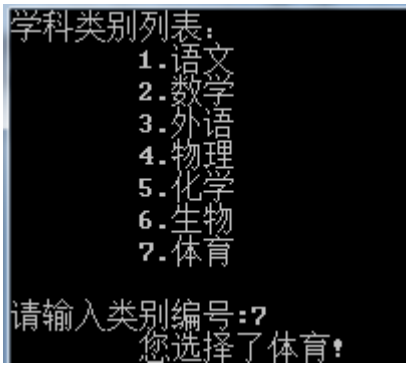
（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

3.试题编号：1-3 任务实现3

（1）任务描述

任务一：请模拟实现一个课程菜单选择功能，如下图所示。(20分)



要求：使用switch语句实现。

任务二：输入一个百分制的成绩t，将其转换成对应的等级然后输出，具体转换规则如下：90~100为A80~；89为B；70~79为C；60~69为D0~59为E。(30分)

要求：使用多分支条件语句实现。

任务三：请完成以下编程工作：

①定义一个动物抽象类Animal，有动物名称，动物打招呼的方法。

②定义它的两个子类Dog和Cat，该类继承动物类。③分别实现它们打招呼的方式。(30分)

要求：

①使用抽象类和抽象方法。

②使用类的继承。

③在主函数(或主方法)中实例化对象，并让对象实现操作。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

4.试题编号：1-4 任务实现4

（1）任务描述

任务一：验证用户输入的数字是否在25-50范围内，如果输入错误或不在25-50范围内就要求用户重新输入。(20分)

要求：利用循环结构完成。

任务二：用户可以无限次的输入数字，请统计用户输入的数字中正数的个数，负数的个数，0的个数。直到用户输入999就结束程序，输出统计结果。(30分)

要求：使用循环结构语句实现。

任务三：求n的阶乘，如果输入的数不在范围之内则要求用户重新输入。(30分)

要求：使用循环结构语句实现。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

5.试题编号：1-5 任务实现5

（1）任务描述

任务一：编写程序实现：商店卖西瓜，20斤以上的每斤0.85元；重于15斤轻于等于20斤的，每斤0.90元；重于10斤轻于等于15斤的，每斤0.95元；重于5斤轻于等于10斤的，每斤1.00元；轻于或等于5斤的，每斤1.05元。输入西瓜的重量和顾客所付钱数，输出应付货款和应找钱数。(20分)

要求：使用分支结构语句实现。

任务二：已知xyz+yzz=532，其中x、y、z均为一位数，编写一个程序求出x、y、z分别代表什么数字。(30分)

要求：使用分支、循环结构语句实现。

任务三：从键盘输入一个整数N，打印出有N*2-1行的菱形。（30分）

```
   *
  ***
 *****
*******
 *****
  ***
   *
```

例如输入整数4，则屏幕输出如下菱形。

现要求输入整数为7，在屏幕中输出相应的菱形。要求：用循环结构语句实现。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

6.试题编号：1-6 任务实现6

（1）任务描述

任务一：有1、2、3三个数字，能组成哪些互不相同且无重复数字的三位数。(20分)

要求：使用循环结构语句实现。

任务二：输入10个学生的单科成绩，求出其中的最高分、最低分、平均分以及超过平均分的人数(30分)

要求：使用数组定义实现。

任务三：使用循环语句打印出如下图案。(30分)

```
*******
```

*****

***

*

要求：使用循环结构语句实现。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

7.试题编号：1-7 任务实现7

（1）任务描述

任务一：根据如下要求计算机票优惠率，并输出。(20分)

输入：用户依次输入月份和需要订购机票的数量，分别保存到整数变量month和sum中。

计算规则如下：

航空公司规定在旅游的旺季7～9月份，如果订票数超过20张，票价优惠15%，20张以下，优惠5%；在旅游的淡季1～5月份、10月份、11月份，如果订票数超过20张，票价优惠30%，20张以下，优惠20%；其他情况一律优惠10%。

输出：根据输入月份和需要订购机票的数量，输出优惠率。

要求：使用分支结构实现上述程序功能。

任务二：使用冒泡排序法对数组中的整数按升序进行排序，如下所示：

原始数组：a[]={1,9,3,7,4,2,5,0,6,8}排序后：a[]={0,1,2,3,4,5,6,7,8,9}(30分)

要求：综合使用分支、循环结构语句实现，直接输出结果不计分。

任务三：输入一个年度，判断是否是闰年。例如，2000是闰年，1900不是闰年，1904是闰年。(30分)

要求：使用分支结构语句实现。

提示：以下两个条件，只要满足任意一个，即是闰年：①能整除4且不能整除100；②能整除400。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

8.试题编号：1-8 任务实现8

（1）任务描述

任务一：输出杨辉三角形，如下图所示：(20分)

*

***

*****

****

****

****

****

要求：使用循环结构语句实现，直接输出结果不计分。

任务二：编程实现判断一个字符串是否是“回文串”。所谓“回文串”是指一个字符串的第一位与最后一位相同，第二位与倒数第二位相同。例如：“159951”、“19891”是回文串，而“2011”不是。(30分)

要求：用带有一个输入参数的函数(或方法)实现，返回值类型为布尔类型。

任务三：任意输入十个数据，打印出改十个数据最大值、最小值。(30分)

要求：①定义一个大小为10的整形数组a；

②从键盘输入10个整数，放置到数组a中；

③输出数组a中的最大值、最小值。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

9.试题编号：1-9 任务实现9

（1）任务描述

任务一：输入n(n<100)个整数，找出其中最小的数，将它与最先输入的数交换后输出这些数。(20分)

要求：用数组解决任务。

任务二：从键盘输入三条边A，B，C的边长，请编程判断能否组成一个三角形。(30分)

要求：A,B,C<1000，如果三条边长A,B,C能组成三角形的话，输出YES，否则NO。

任务三：对于给定的一个字符串，统计其中数字字符出现的次数。(30分)

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

10.试题编号：1-10 任务实现10

（1）任务描述

任务一：某运输队为超市运送暖瓶500箱，每箱装有6个暖瓶。已知每10个暖瓶的运费为5元，损坏一个不但不给运费还要赔10元，运后结算时，运输队共得1350元的运费。编程输出损坏暖瓶的个数。(20分)

要求：用循环语句实现

任务二：编写程序，从键盘接收一个只包含英文字母的字符串，对字符串中的字母进行大小写互转（大写字母转成小写，小写字母转成大写）。（30分）

例如键盘输入：abcABC输出：ABCabc

要求：使用循环和判断语句实现。

任务三：对于给定的一个字符串，统计其中数字字符出现的次数。（30分）

要求： 字符串只能由数字和字符组成。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

程序设计模块附录

附录1作品提交

①请建立以“考生号_题号”命名的成果文件夹，所有提交文件均放在该目录下。例如：144115040001_T1_1；

②分别将每个任务的代码以成员函数的形式封装到类中，并且在main函数中调用该成员函数；

③在成果文件夹中创建三个文件夹task1、task2、task3，将三个任务的源代码、编译后的文件及对应成员函数的程序流程图截图分别保存至相应文件夹；

④将成果文件夹压缩打包，按照要求上传至服务器。

⑤考核时间为180分钟。

附录2实施条件

表1考点提供的主要设备及软件表

序号	场地、设备、软件名称	规格/技术参数、用途	备注
1	大数据技术实训机房	测试场地	保证参考人员有足够间距
2	计算机	CPU酷睿i5以上，内存4G以上，win7/win10/linux操作系统	用于软件开发和软件部署，每人一台
3	Pycharm2018.2以上、IntelliJ IDEA 2018.2以上、Eclipse4.7或以上	软件开发	参考人员自选一种开发工具
4	MSDN或者JDK帮助文档中文版	帮助文档	参考人员可以使用帮助文档

附录3评价标准

评分标准一：实操文档（10分）

表2 实操文档评分细则表

序号	评分项	分值	评分细则
1	实操文档有无	2分	有实操文档得分，无实操文档扣2分。
2	文档任务截图	4分	有操作过程截图得分，无操作过程截图扣4分。
3	文档任务截图标注	4分	有文档任务截图标注说明和画框得分，无标注和画框扣4分。

评分标准二：依据题的任务，完成任务功能（80分）

表3 项目功能评分细则表

序号	评分项	分值	评分细则
1	任务实现	80分	试题按任务分值评分;未按要求提交正确格式的源文件，扣5分;程序中出现了没有使用的变量扣1分；程序中出现了无用的循环、分支、循序结构扣1分，扣完为止。

评分标准三：职业素质（10分）

表4 职业素质评分细则表

序号	评分项	分值	评分细则
1	代码书写格式规范	3分	代码缩进不规范扣1分、方法划分不规范扣1分、语句结构不规范扣1分（如一行编写两个语句）、使用空行不规范扣1分，扣完为止。
2	注释规范	2分	整个项目没有注释扣2分、有注释，但注释不规范扣1分，扣完为止。
3	类名、变量名、方法名命名规范	5分	命名规范，为满分。类名、变量名或方法名命名不规范或没有实际意义的每个扣1分，扣完为止。

模块二 数据库设计与开发

1.试题编号：2-1《教务管理系统》项目教材订购管理模块

（1）任务描述

《教材订购管理》模块的 E-R图如图2.1.1 所示，逻辑数据模型如图2.1.2所示，物理数据模型如图 2.1.3所示，数据表字段名定义见表2.1.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

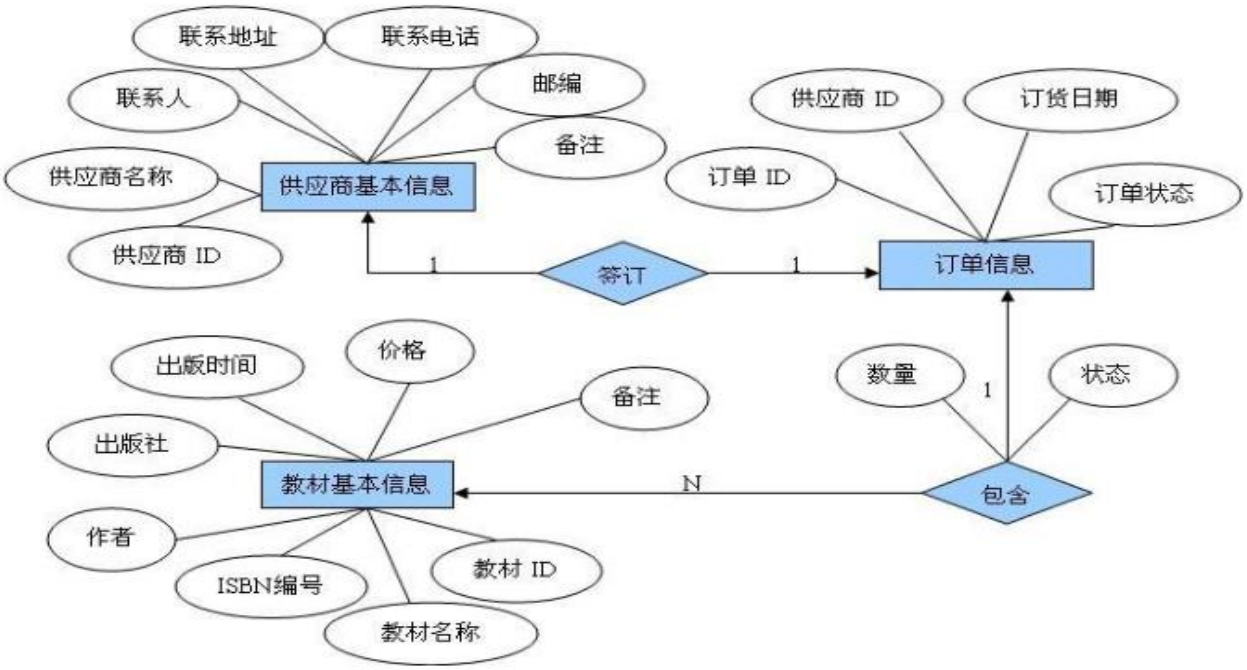


图2.1.1E-R图

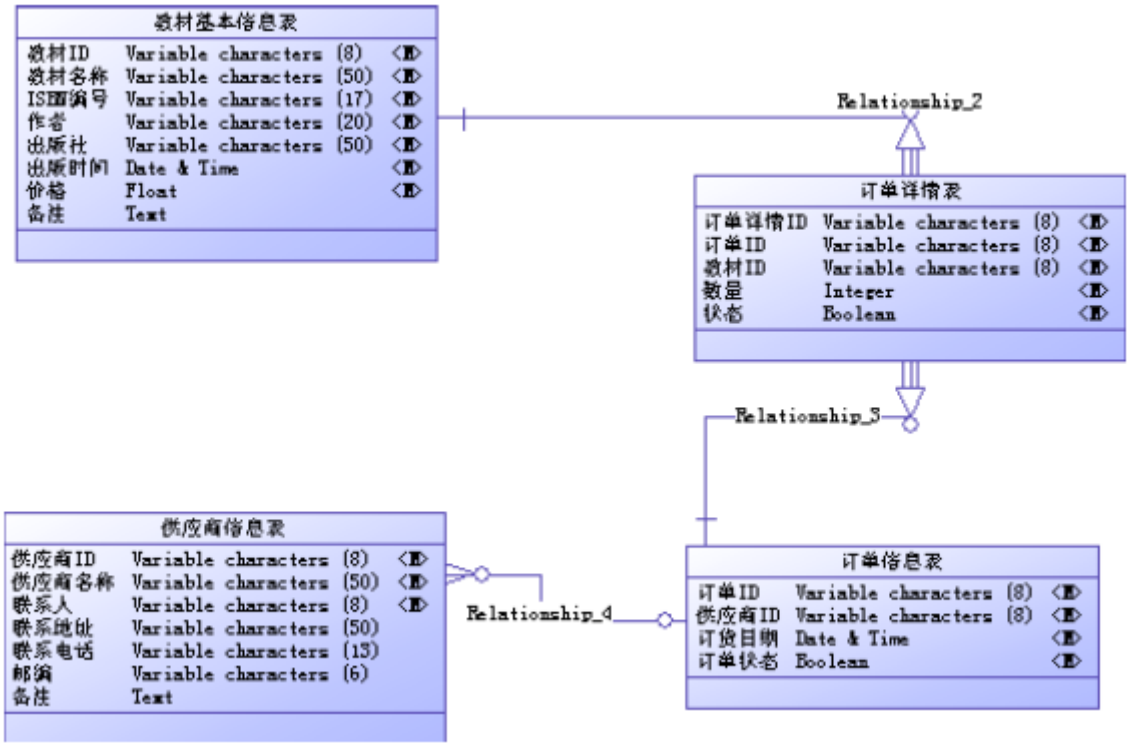


图2.1.2逻辑数据模型图

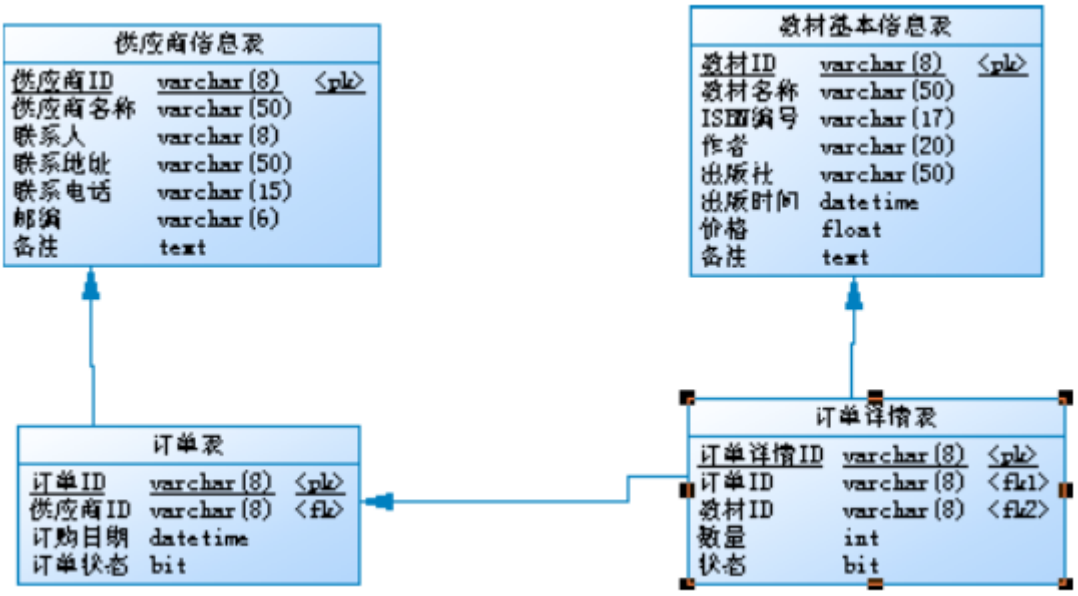


图2.1.3物理数据原型图

表2.1.1字段名定义表

字段名	字段说明	字段名	字段说明
book_id	教材 id	supplier_name	供应商名称
book_name	教材名称	supplier_people	联系人
book_isbn	教材 ISBN 编号	supplier_address	联系地址
book_author	作者	supplier_phone	联系电话
book_publisher	出版社	supplier_postcode	邮编
book_price	价格	supplier_remark	备注
book_rkm	备注	orderdet_id	订单详情 id
order_id	订单 id	orderdet_status	订单详情状态
order_datetime	订购时间	book_datetime	出版时间
order_status	订单状态	orderdet_num	数量
supplier_id	供应商 id		

任务一：创建数据库（10 分）

创建数据库HNIUEAM。

任务二：创建数据表（25 分）

根据图 2.1.2 和表 2.1.1，创建数据表 T_Supplier、 T_BookInfo、 T_Order。

任务三：创建数据表间的关系及约束（15 分）

根据物理数据原型，创建数据关系表。

任务四：数据操作（25 分）

用SQL语句完成如下操作：

①.向T_Supplier表插入数据：“BC0001,windows程序设计,0257-9413,刘立,电子工业出版社代理商,2010-11-10,42,无”；

②.查询出供应商名称为“电子工业出版社代理商”的订单编号及订单状态；

③.查询教材名称为“windows程序设计”的订购日期；

④.创建视图查询供应商名为“电子工业出版社代理商”所订购的教材的详细信息(包括名称，ISBN编号，

作者，出版社，出版时间，价格，数量)；

⑤.创建存储过程，当订单详情表中相应订单的状态为“1”时,修改订单表的订单状态为“1”。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

2.试题编号：2-2《图书管理信息系统》 项目

（1）任务描述

《图书管理信息系统》中借书管理子模块的 E-R 图如图 2.2.1 所示，逻辑数据模型如图2.2.2所示，物理数据模型如图2.2.3所示，数据表字段名定义见表2.2.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

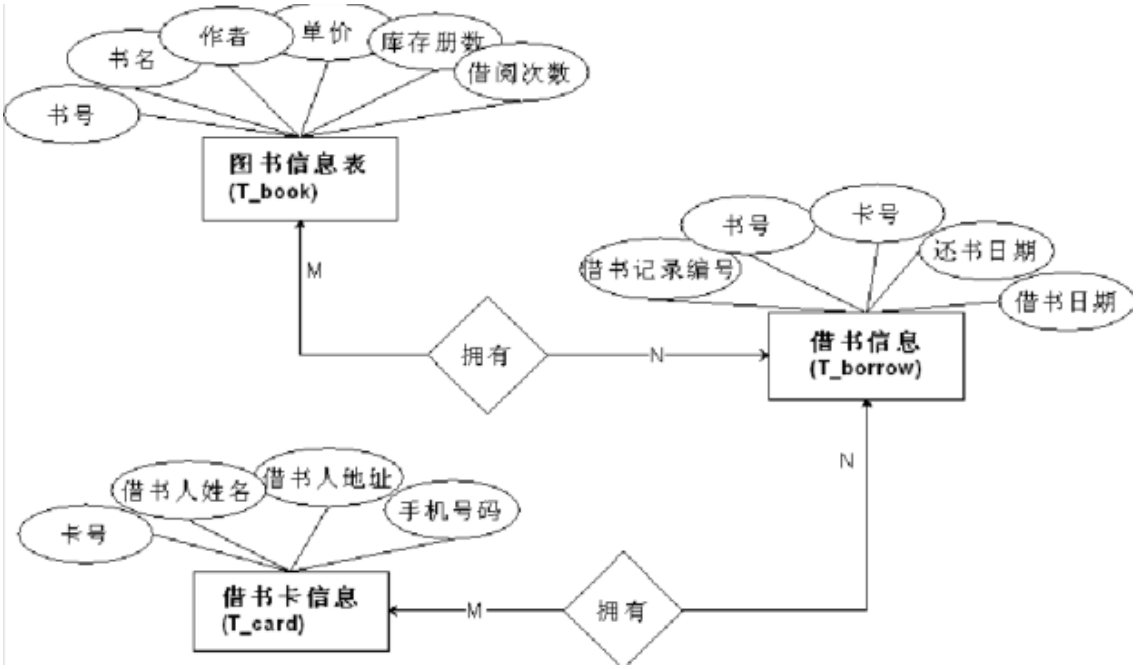


图2.2.1E-R图

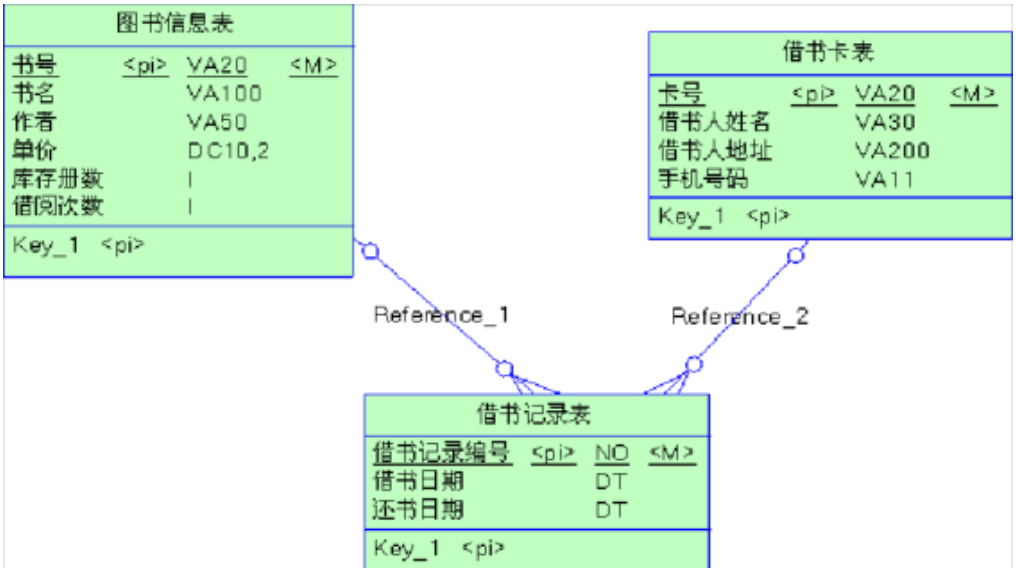


图2.2.2逻辑数据模型图

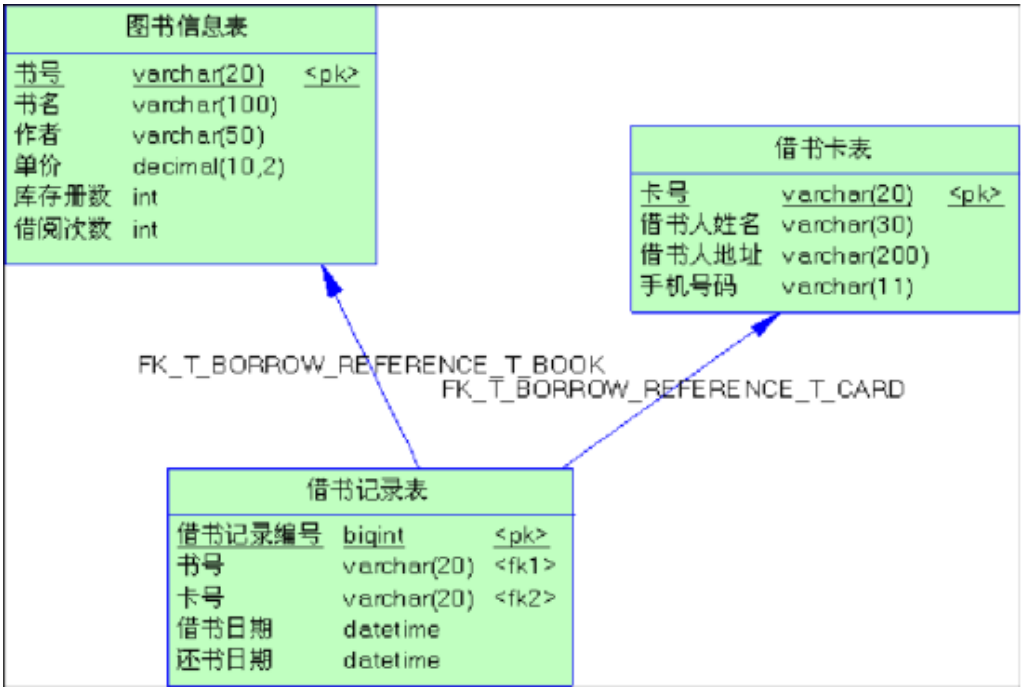


图2.2.3物理数据原型图

表2.2.1字段名定义表

字段名	字段说明	字段名	字段说明
Book_no	书号	Card_name	借书人姓名
Book_name	书名	Adress	借书人地址
Author	作者	Mobile	手机号码
Price	单价	Borrow_id	借书记录编号
Qty	库存册数	Borrow_date	借书日期
Loan_qty	借阅次数	Return_date	还书日期
Card_no	卡号		

任务一：创建数据库（10 分）

创建数据库 BookDB。

任务二：创建数据表（25 分）

根据图2.2.2和表2.2.1，创建数据表T_card、T_book、T_borrow。

任务三：创建数据表间的关系及约束（15 分）

根据物理数据原型，创建数据关系表。

任务四：数据操作（25 分）

用SQL语句完成如下操作：

①.向每个表插入3条测试数据；

②.将“李”姓作者的所有图书单价下调10%；

③.查询出日期在2010-10-31至2010-11-31之间借出的图书信息；

④.查询出手机号为“135”开头的所有借书人姓名；

⑤.创建视图查询库存数量小于10册的图书信息；

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

3.试题编号：2-3《学生管理信息系统》项目

（1）任务描述

《学生管理信息系统》中成绩管理子模块的E-R 图如图2.3.1所示，逻辑数据模型如图2.3.2所示， 物理数据模型如图2.3.3所示，数据表字段名定义见表2.3.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

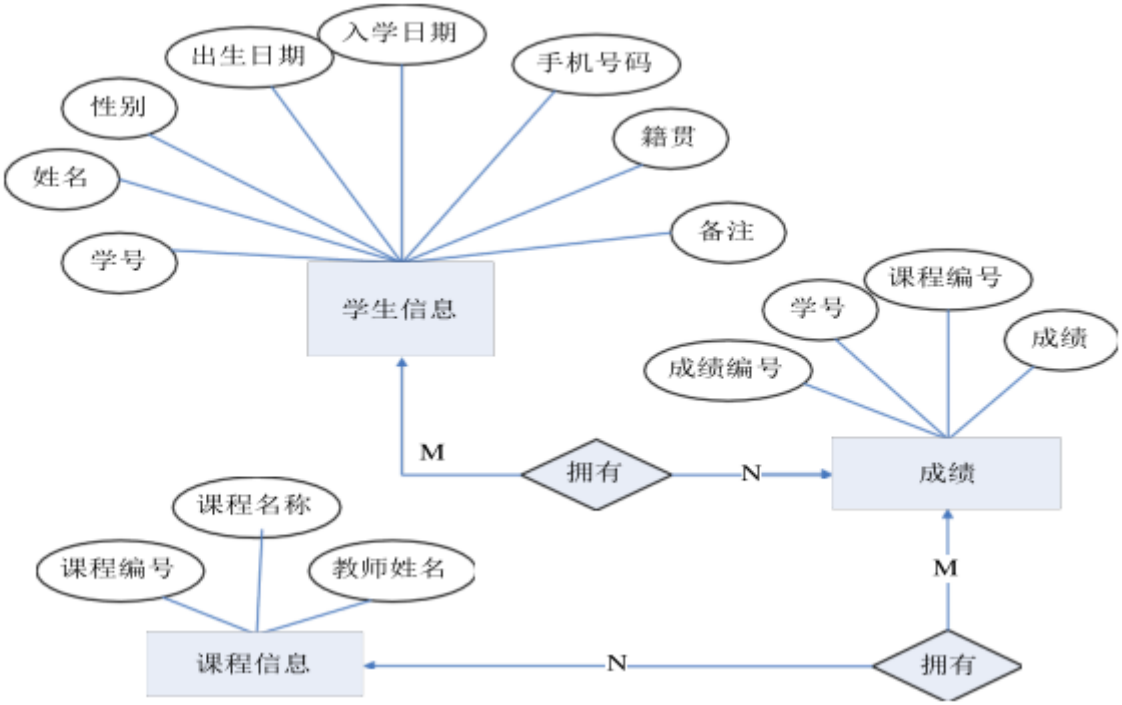


图2.3.1E-R图

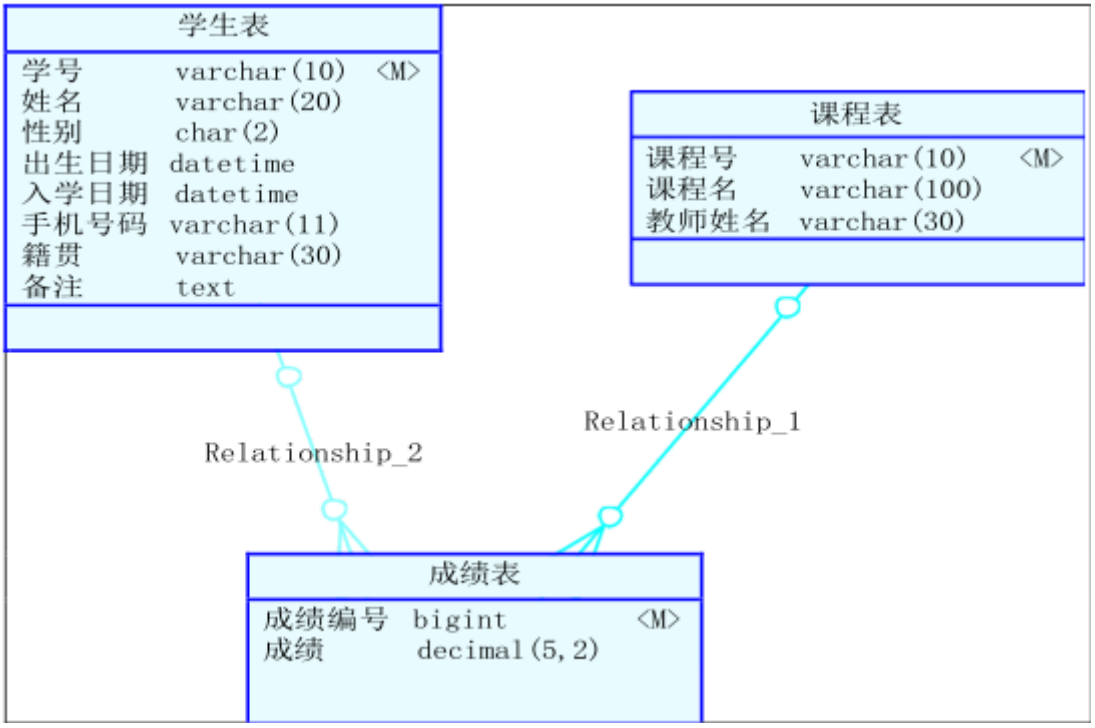


图2.3.2逻辑数据模型图

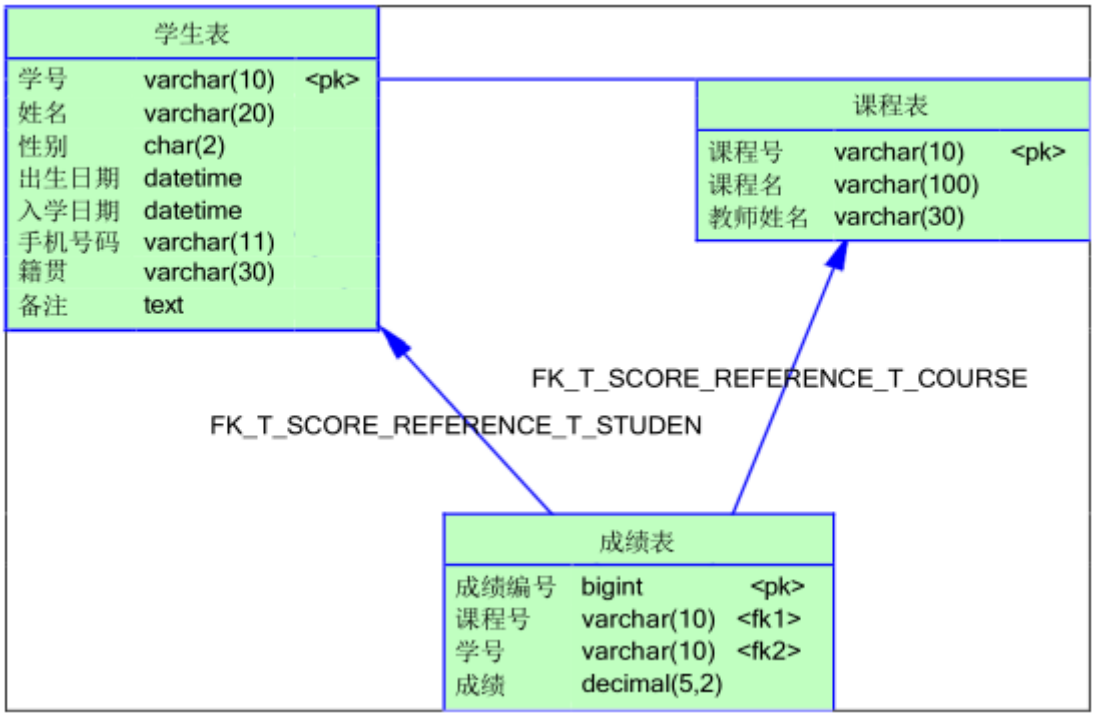


图2.3.3物理数据模型图

表2.3.1字段名定义表

字段名	字段说明	字段名	字段说明
Stud_id	学号	Reserve	备注
Stud_name	姓名	Course_id	课程编号
Stud_sex	性别	Course_name	课程名称
Birth_date	出生日期	Teacher_name	教师姓名
Entry_Date	入学日期	Score_id	成绩编号
Mobile	手机号码	Score	成绩
Birth_place	籍贯		

任务一：创建数据库（10 分）

创建数据库 StudentDB。

任务二：创建数据表（25 分）

根据图2.3.2 和表2.3.1，创建数据表T_student、T_course、T_score。

任务三：创建数据表间的关系及约束（15 分）

根据物理数据原型，创建数据关系。

任务四：数据操作（25 分）

用SQL语句完成如下操作：

①.向每个表插入3条测试数据；

②.删除所有选修“日语”的同学的选课记录；

③.查询出“数据库原理” 这门课的最高成绩；

④.查询出所有选修了“数据库原理”课程的学生学号、姓名和籍贯；

⑤.创建视图，查询指定课程名称的平均成绩。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

4.试题编号：2-4《人力资源管理系统》项目

（1）任务描述

《人力资源管理系统》中人员管理子模块的 E-R 图如图2.4.1 所示，逻辑数据模型如图2.4.2所示，物理数据模型如图2.4.3所示，数据表字段名定义见表2.4.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

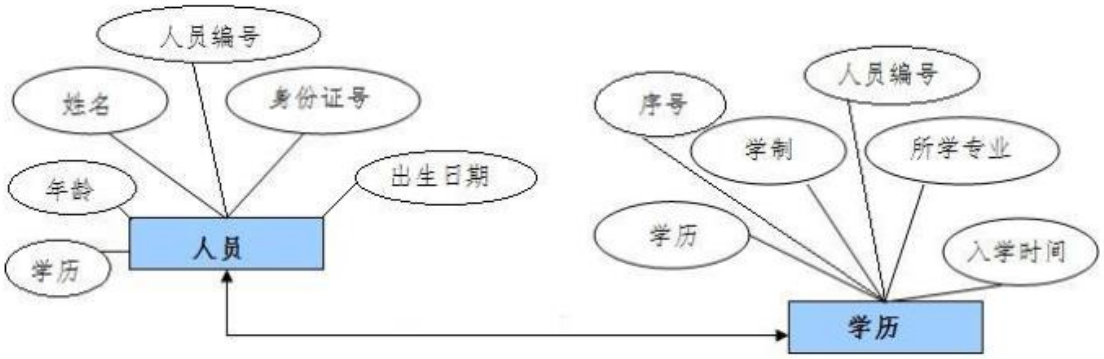


图2.4.1E-R图

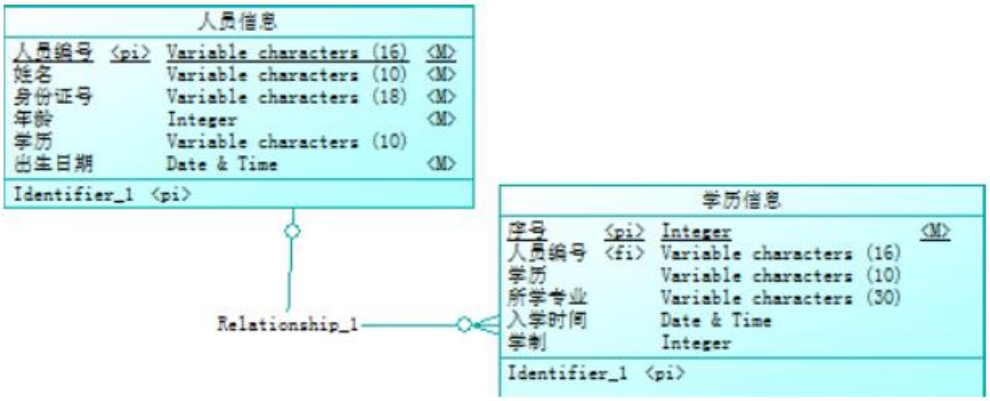


图2.4.2逻辑数据模型图

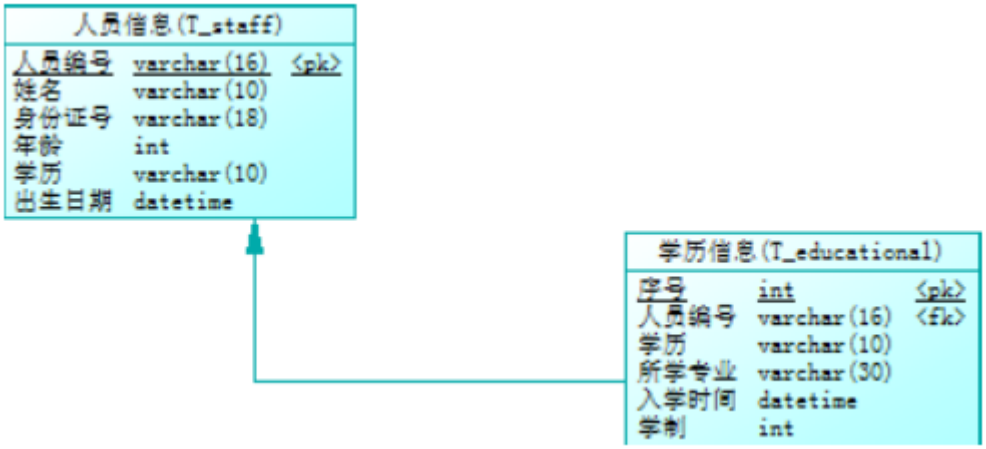


图2.4.3物理数据模型图

表2.4.1字段名定义表

字段名	字段说明	字段名	字段说明
staff_no	人员编号	id	序号(自动增长)
name	姓名	degree	学历
ic_card	身份证号	major	所学专业
age	年龄	reg_time	入学时间
bithday	出生日期	length_of_schooling	学制

任务一：创建数据库（10 分）

创建数据库ResourcesDB。

任务二：创建数据表（25 分）

根据图2.4.2 和表 2.4.1，创建数据表T_staff、T_educational。

任务三：创建数据表间的关系及约束（15 分）

①.为表设置主键，主键命名为“pk_<表名>_<主键标识>”；

②.根据逻辑数据模型，创建数据表之间的关系，关系命名为“fk_<表名>_<主表名>_<外键标识>”；

任务四：数据操作（25 分）

用SQL语句完成如下操作：

①.向每个表插入2条测试数据；

②.查询出T_staff表中大于平均年龄的人员名单；

③.查询出入学时间在 2015-9-1 之后的所有人员名单；

④.查询出学习“大数据技术”专业的所有人员名单；

⑤.创建存储过程，根据入学时间和学制计算每个人的毕业年份数。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

5.试题编号：2-5《员工工资管理》项目

（1）任务描述

《员工工资管理》的 E-R图如图2.5.1所示，逻辑数据模型如图2.5.2所示，物理数据模型如图 2.5.3所示，数据表字段名定义见表2.5.1 。请按以下设计完成数据库创建、数据表创建和数据操作任务：

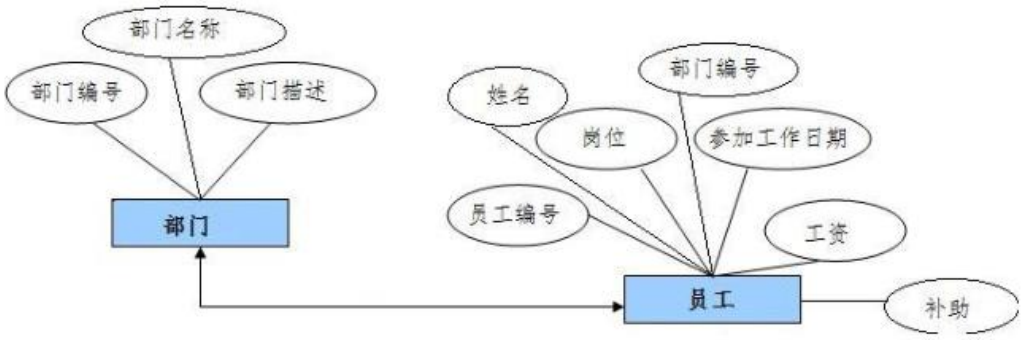


图2.5.1E-R图

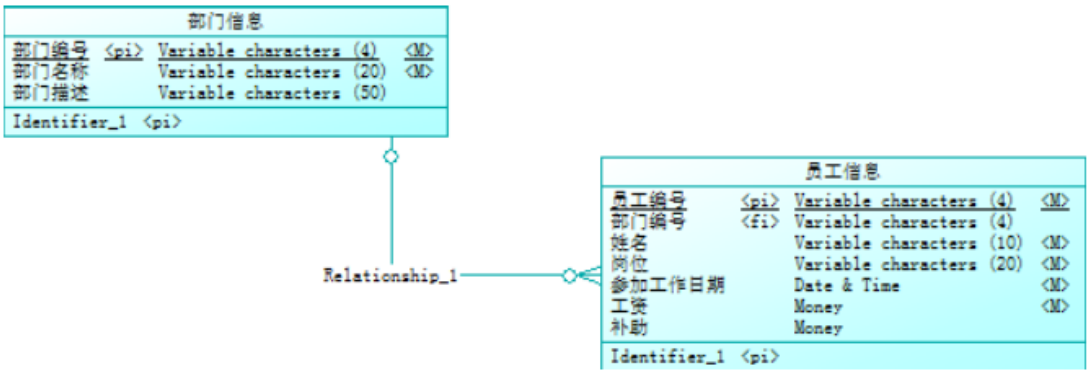


图2.5.2逻辑数据模型图

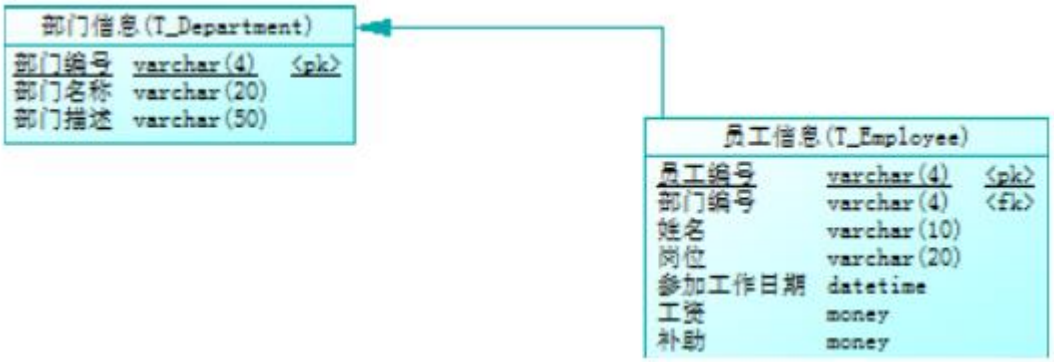


图2.5.3物理数据模型图

表2.5.1字段名定义表

字段名	字段说明	字段名	字段说明
dep_no	部门编号	post	岗位
dep_name	部门名称	work_time	参加工作日期
dep_desc	部门描述	salary	工资
emp_no	员工编号	bonus	补助
name	姓名		

任务一：创建数据库（10 分）

创建数据库SalaryDB。

任务二：创建数据表（25 分）

根据图2.24.2 和表2.24.1，创建数据表T_Department、T_Employee。

任务三：创建数据表间的关系及约束（15 分）

①.创建主键（两个表均设置）；

②.根据逻辑数据模型，创建数据表之间的关系，关系命名为“fk_<表名>_<主表名>_<外键标识>”；

任务四：数据操作（25 分）

用SQL语句完成如下操作：

①.向每个表插入2条测试数据；

②.查询出所有已有的岗位，要求取出重复项；

③.查询出每个部门每种岗位的平均工资和最高工资。

④.创建视图，显示所有没有补助的员工的姓名；

⑤.创建存储过程，显示平均工资低于3500的部门编号、平均工资、最高工资，要求以平均工资升序排序。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

6.试题编号：2-6《自学考试网》 项目

（1）任务描述

《自学考试网》的E-R 图如图2.6.1所示，逻辑数据模型如图2.6.2所示，物理数据模型如图2.6.3所示，数据表字段名定义见表2.6.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

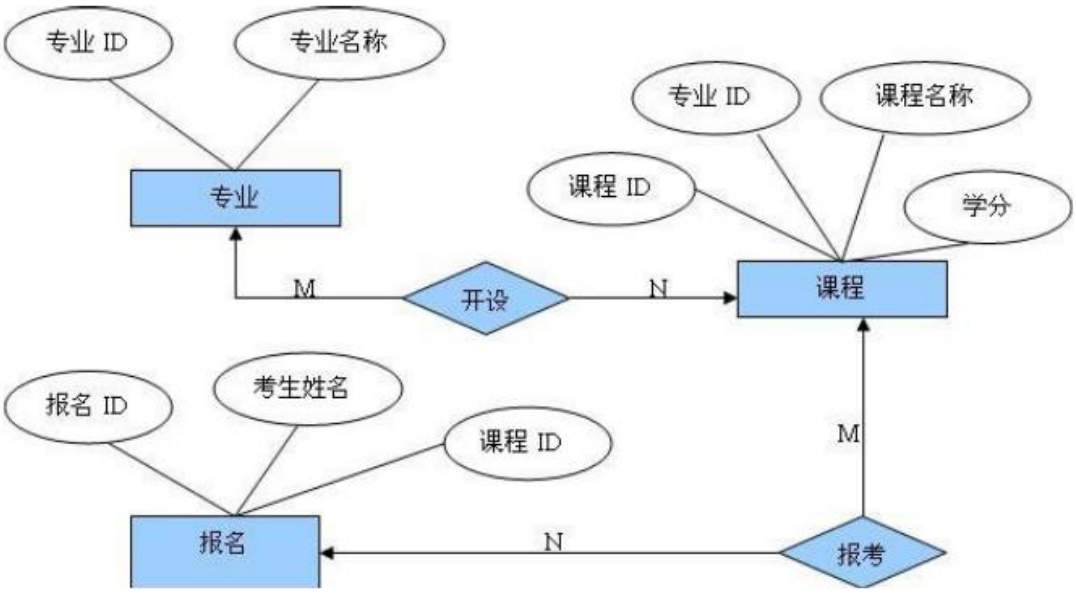


图2.6.1E-R图

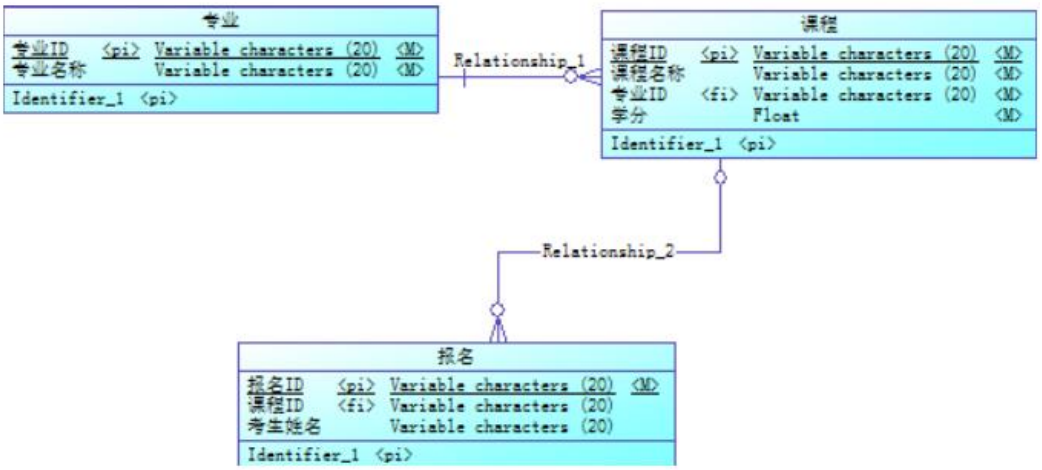


图2.6.2逻辑数据模型图

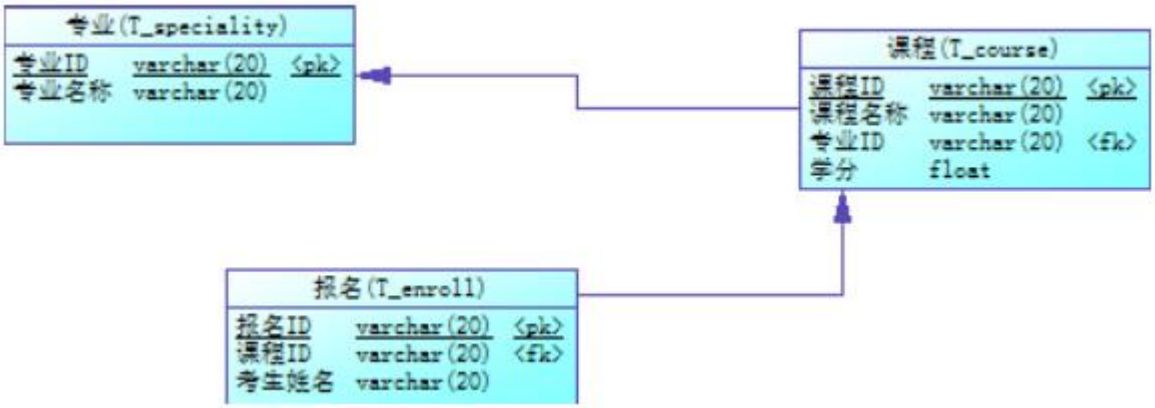


图2.6.3物理数据模型图

表2.6.1字段名定义表

字段名	字段说明	字段名	字段说明
id<pk>	专业 ID	mark	课程学分
name	专业名称	id<pk>	报名 ID
id<pk>	课程 ID	course_id	课程 ID
specialityid	专业 ID	name	考生姓名
name	课程名称		

任务一：创建数据库（10 分）

创建数据库SelfStudy。

任务二：创建数据表（25 分）

根据图2.6.2和表2.6.1，创建数据表T_speciality、T_course、T_enroll。

任务三：创建数据表间的关系及约束（15 分）

①.创建主键（三个表均设置）；

②.创建数据表之间的关系，关系命名为“fk_<表名>_<主表名>_<外键标识>”；

任务四：数据操作（25 分）

利用数据管理工具在表中插入以下数据， 用作测试。

表2.6.2 T_speciality表测试数据

id	name
001	大数据技术
002	计算机网络技术
003	人工智能技术应用

表2.6.3 T_course表测试数据

id	speciality_id	name	mark
001	001	Python程序设计	3
002	001	网页设计	3
003	001	爬虫应用技术与开发	3

表2.6.4 T_enroll表测试数据

id	course_id	name
001	001	王明
002	002	王明
003	003	王明

用SQL语句完成如下操作：

①.在T_course表插入数据：“004，001，高等数学，3”；

②.查询“大数据技术”专业开设的课程；

③.查询“大数据技术”专业有哪些考生报名；

④.查询出报考课程为“网页制作”的考生；

⑤.创建可查询考生姓名，报考课程名称的视图；

⑥.创建存储过程，查询报考某门课程（以课程名称为参数）的考生。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

7.试题编号：2-7《图书管理信息系统》项目

（1）任务描述

《图书管理信息系统》的 E-R 图如图2.7.1 所示，逻辑数据模型如图 2.7.2 所示，物理数据模型如图2.7.3 所示，数据表字段名定义见表2.7.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

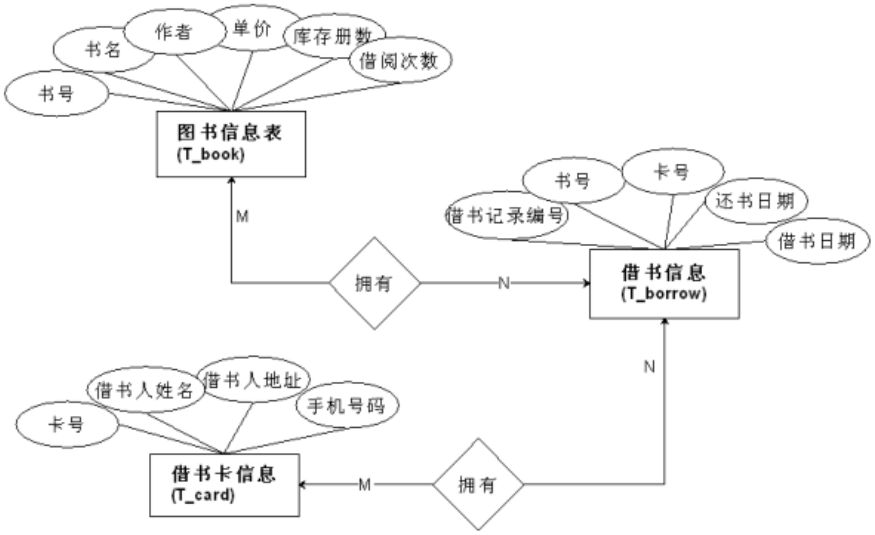


图2.7.1E-R图

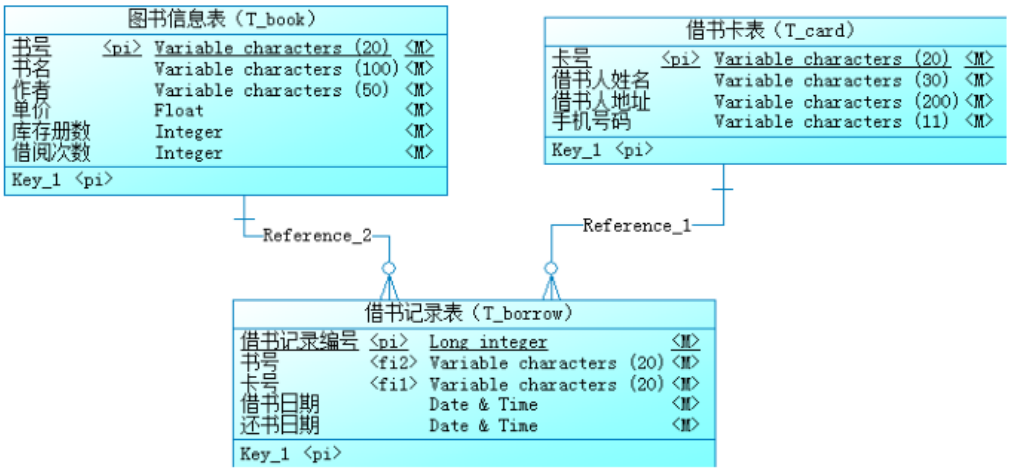


图2.7.2逻辑数据模型图

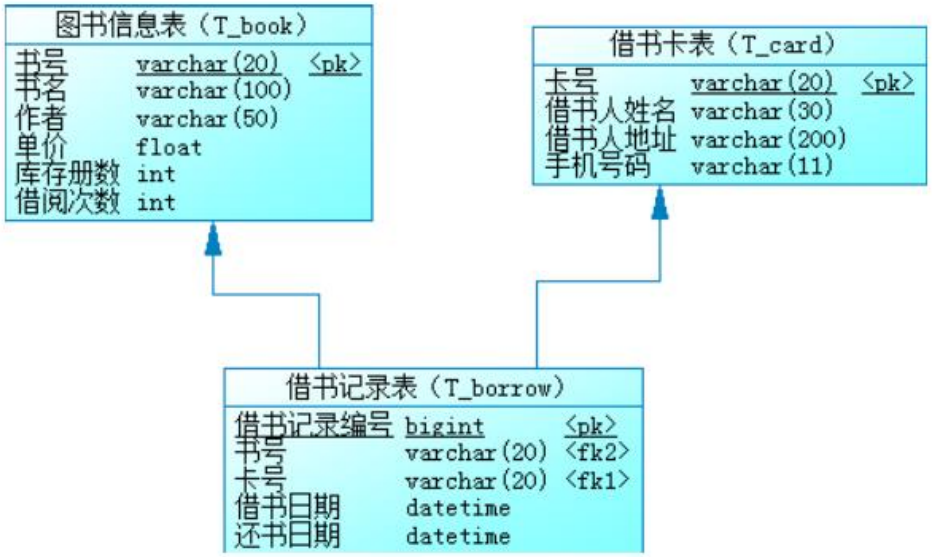


图2.7.3物理数据模型图

表2.7.1字段名定义表

字段名	字段说明	字段名	字段说明
book_no	书号	card_name	借书人姓名
book_name	书名	adress	借书人地址
author	作者	mobile	手机号码
price	单价	borrow_id	借书记录编号
qty	库存册数	borrow_date	借书日期
loan_qty	借阅次数	return_date	还书日期
card_no	卡号		

任务一：创建数据库（10 分）

创建数据库 BookDB。

任务二：创建数据表（25 分）

根据图2.7.2 和表 2.7.1，创建数据表T_card、T_book、T_borrow。

任务三：创建数据表间的关系及约束（15 分）

根据物理数据原型，创建数据关系。

任务四：数据操作（25 分）

用SQL语句查询出如下数据：

①.在T_book 表中插入数据：“9787302245339，Access数据库技术与应用，陈世红，27.20，50”；

②.查询出日期为2010-10-31以后借出的图书信息；

③.查询出没有还书的借书人姓名；

④.创建视图查询借书人的姓名，手机号码和地址；

⑤.查询出库存数量小于5册的图书信息；

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

8.试题编号：2-8《银行信贷管理系统》项目

（1）任务描述

《银行信贷管理系统》的E-R 图如图2.8.1 所示，逻辑数据模型、物理数据模型如图2.8.2和图2.8.3所示。 数据表字段名定义见表 2.8.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

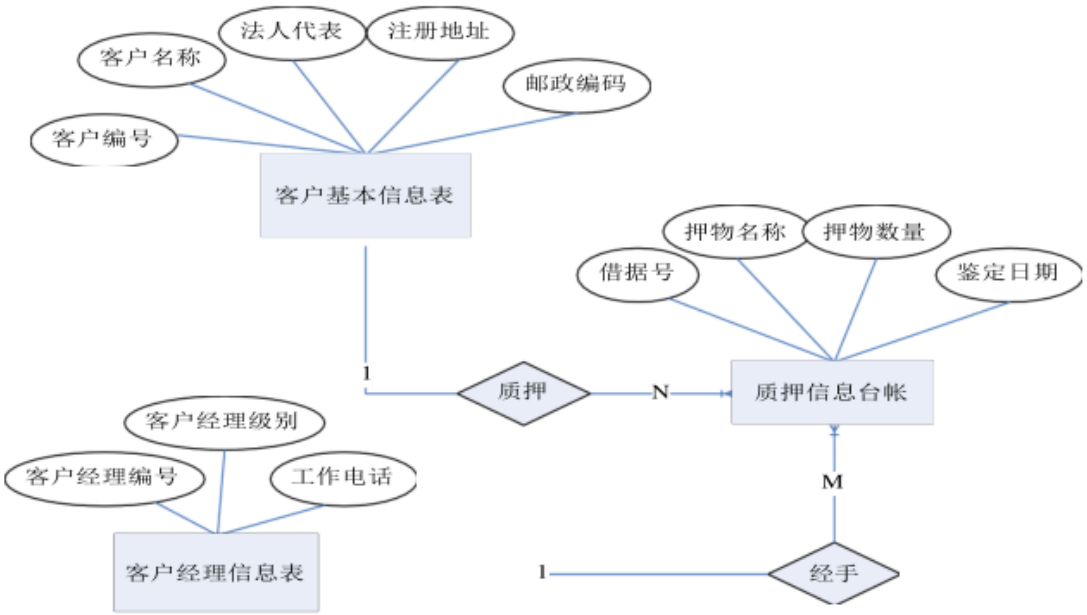


图2.8.1E-R图

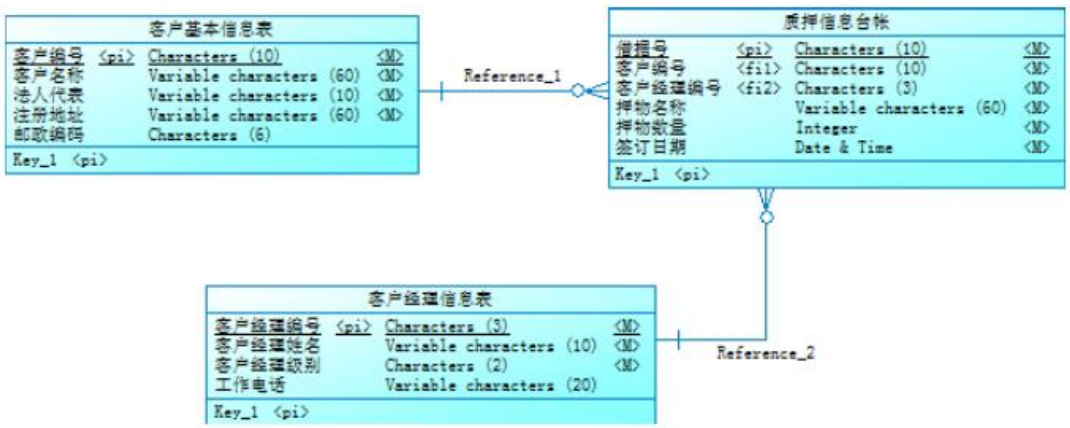


图2.8.2逻辑数据模型图

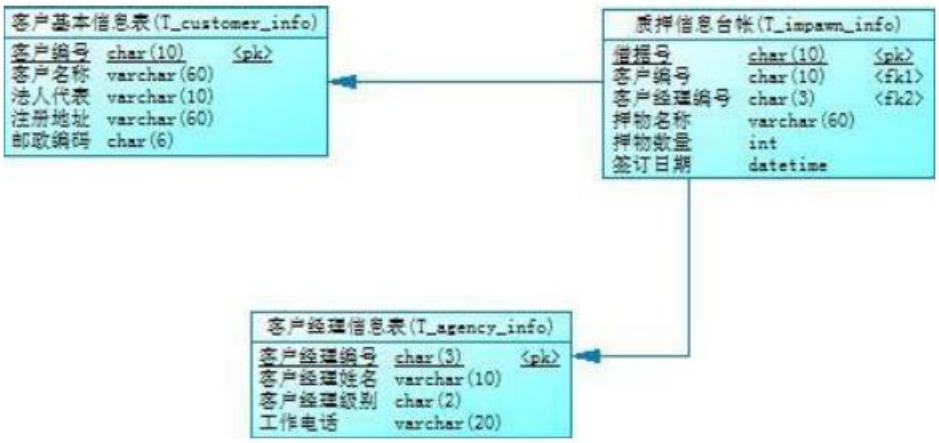


图2.8.3物理数据模型图

表2.8.1字段名定义表

字段名	字段说明	字段名	字段说明
agency_id	客户经理编号	reg_address	注册地址
agency_name	客户经理姓名	post_code	邮政编码
agency_level	客户经理级别	borrow_id	借据号
cust_id	客户编号	pawn_goods_name	押物名称
cust_name	客户名称	pawn_goods_num	押物数量
legal_name	法人代表	contract_date	签订日期
agency_phone	工作电话		

任务一：创建数据库（10分）

创建数据库 BankCreditLoanDB。

任务二：创建数据表（25分）

根据图2.4.2和表2.4.1，创建数据表T_customer_info、T_impawn_info、T_agency_info。

任务三：创建数据表间的关系及约束（15分）

①.为表设置主键，主键命名为“pk_<表名>_<主键标识>”；

②.根据逻辑数据模型，创建数据表之间的关系，关系命名为“fk_<表名>_<主表名>_<外键标识>”；

③.要求邮政编码由6位数字组成。

任务四：数据操作（25 分）

用SQL语句执行以下操作：

①.分别向三个表中插入一条测试数据，其中客户经理编号为“001”；

②.查询“XX 公司”质押的物品及数量（说明：“XX 公司”为插入测试数据中的公司名称）；

③.统计每个客户经理所经手的质押业务数，查询结果集应包含字段：客户经理姓名、质押业务数；

④.创建存储过程P_customer_info，删除指定客户编号的客户基本信息，同时也删除该客户在质押信息台帐中的所有记录。

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

9.试题编号：2-9《建设工程监管信息系统》项目系统权限管理模块

（1）任务描述

《系统权限管理》模块的 E-R 图如图 2.9.1 所示，逻辑数据模型如图 2.9.2 所示，物理数据模型如图 2.9.3 所示，数据表字段名定义见表 2.9.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

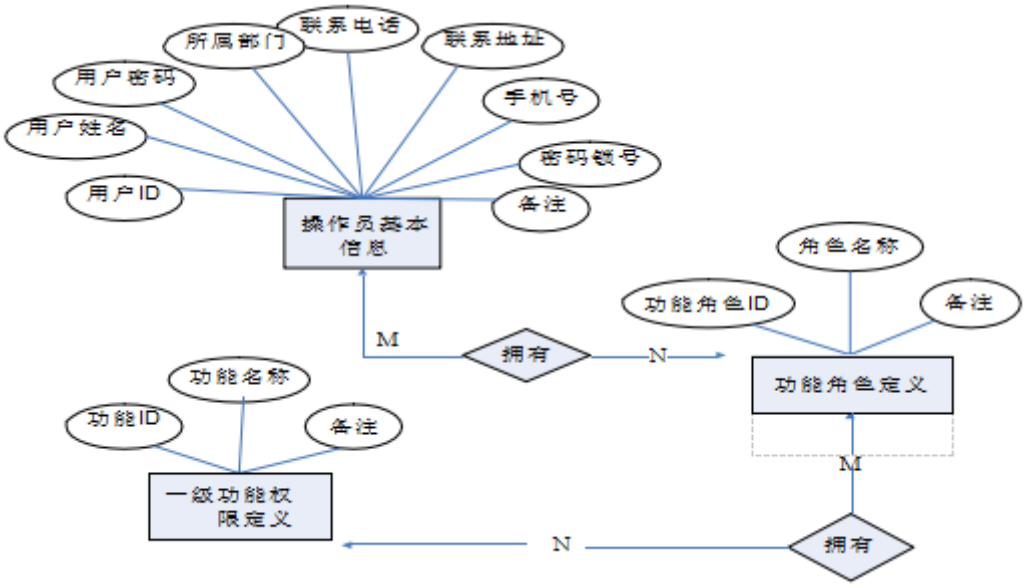


图2.9.1E-R图

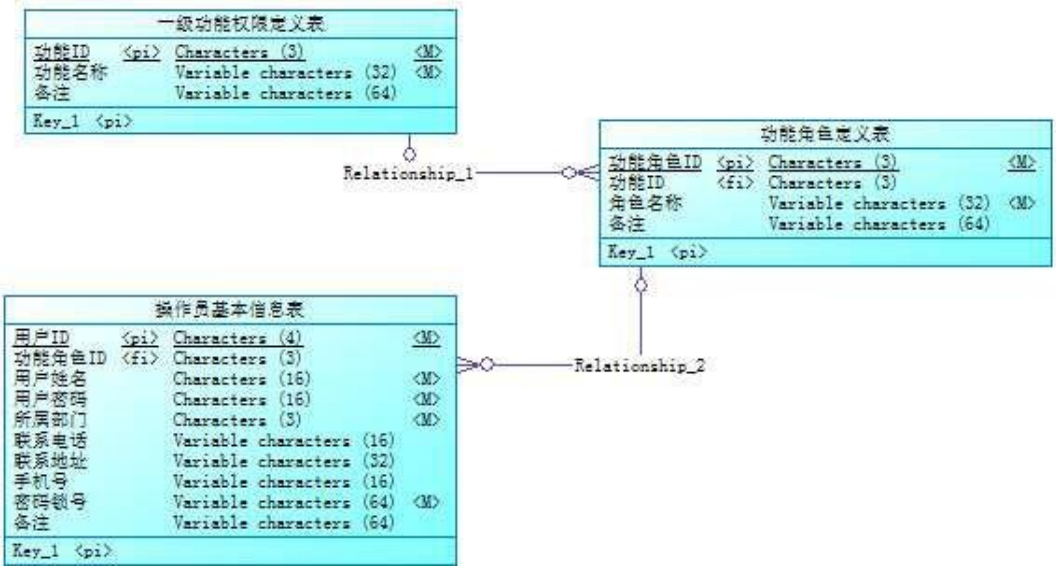


图2.9.2逻辑数据模型图

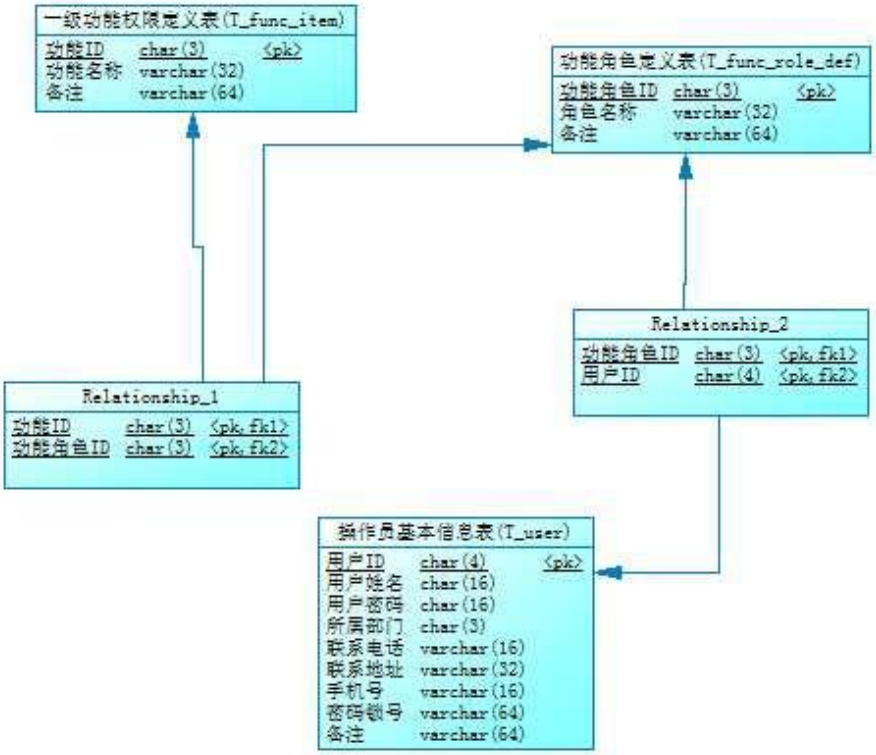


图2.9.3物理数据模型图

表2.9.1字段名定义表

字段名	字段说明	字段名	字段说明
func_id	功能 id	user_passwd	用户密码
func_name	功能名称	dept_id	所属部门
func_role_id	功能角色 id	telephone	联系电话
func_role_name	角色名称	address	联系地址
user_id	用户 id	handphone	手机号
user_name	用户姓名	usb_no	密码锁号
reserve	备注		

任务一：创建数据库（10 分）

创建数据库 ConstructionDB。

任务二：创建数据表（25 分）

根据图 2.1.2 和表 2.1.1，创建数据表 T_user、T_func_item、T_func_role_def 及两个关系表（关系表的名字自拟）。

任务三：创建数据表间的关系及约束（15 分）

根据物理数据原型，创建数据关系表。

任务四：数据操作（25 分）

用 SQL 语句完成如下操作：

①.在T_user 表插入数据“：id01，刘德华，123，KBB，5678900，湖南长沙，13899005678，ldh123，admin”；

②.查询出所属部门为“KBB”的操作员的基本信息；

③.查询出姓名为“刘德华”的操作员具有哪些功能权限；

④.查询出“投标责任人”角色所拥有的功能；

5.创建视图查询操作员的姓名，密码和所属部门；

6.创建存储过程，查询指定操作员所具有的功能权限。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

10. 试题编号：2-10《某电子商务网站》项目产品管理模块

任务描述

《产品管理》模块的 E-R 图如图 2.10.1 所示，逻辑数据模型如图 2.10.2 所示，物理数据模型如图 2.10.3 所示，数据表字段名定义见表 2.10.1。请按以下设计完成数据库创建、数据表创建和数据操作任务：

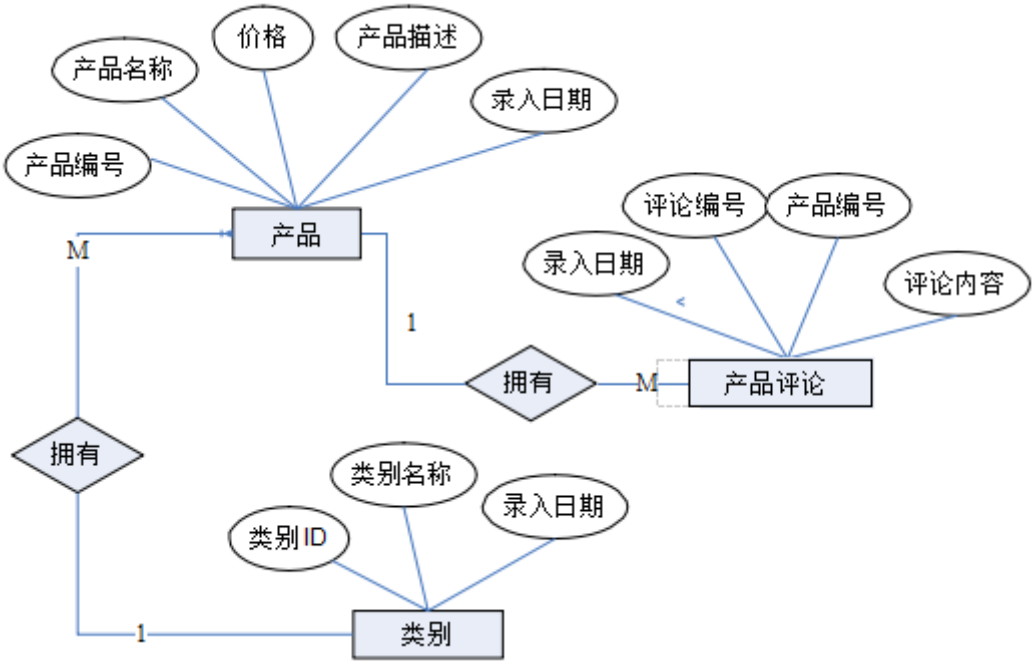


图2.10.1E-R图

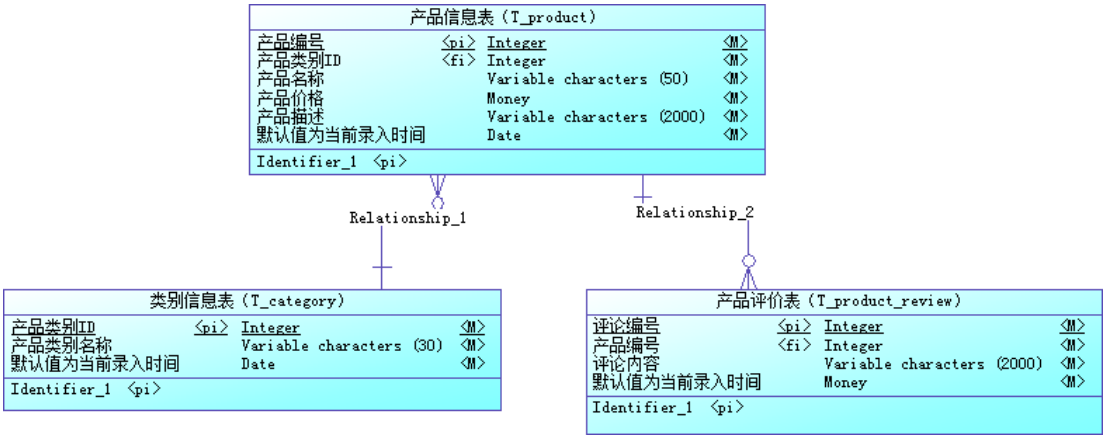


图2.10.2逻辑数据模型图

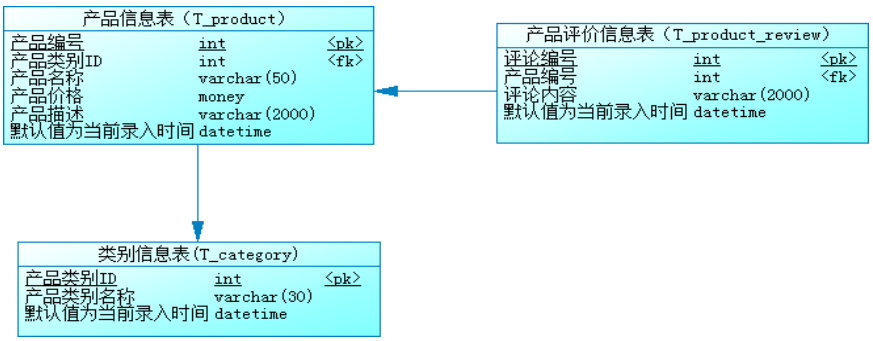


图2.10.3物理数据原型图

表2.10.1字段名定义表

字段名	字段说明	字段名	字段说明
category_id 标识列	产品类别 ID	remark	产品描述
category_name	产品类别名称	register_date	默认值为当前录入时间
register_date	默认值为当前录入时间	review_id 标识列	评论编号
product_id	产品编号	product_id 标识列	产品编号
category_id	产品类别 ID	review	评论内容
product_name	产品名称	register_date	默认值为当前录入时间
price	产品价格		

任务一：创建数据库（10 分）

创建数据库 ProductDB。

任务二：创建数据表（25 分）

根据图2.10.2 和表2.10.1，创建数据表 T_category、T_product_review、T_product，其中产品表的产品ID(product_id)列设置为标识列，自动从 1 开始增长。

任务三：创建数据表间的关系及约束（15 分）

①.创建主键（三个表均设置）；

②.产品价格列（Price）只能输入 1-1000 之间的数；

③.录入时间列(Register_date)默认值为当前录入时间（三个表均设置）。

任务四：数据操作（25 分）

①.用 SQL 语句查询出如下数据：在三个表分别中录入 3 条测试数据（样本数据包含下面题目中使用的数据）；

②.查询某类别下所有产品；

③.查询产品价格在 300-500 元之间的产品；

④.查询录入日期在 2011 年 3 月到 6 月之间的产品数据；

⑤.查询产品价格在 90-200 元之间的所有评论；

⑥.查询评论数在 1-3 条的所有产品。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

数据库设计模块附录

附录1作品提交

答案以“答题文件”的形式提交。请按以下要求创建答题文件夹和答题文件：

①创建答题文件夹

创建以”考生号_题号”命名的文件夹，存放所有答题文件，例如：“340103**********_2_1\”

②创建答题文件

▪ SQL脚本文件

创建project.sql文件，如：“340103**********_2_1\project.sql，存放SQL脚本代码。

▪ 数据库文件

创建db子文件夹，如:“340103**********_2_1\db\”，存放数据库备份文件，它用于教师阅卷时还原数据库。

③提交答题文件

将“考生号_题号”文件夹打包，形成“考生号_题号.rar”文件，如：“340103**********_2_1.rar”，将该文件按要求进行上传。

④考核时量

考核时长为180分钟。

附录2实施条件

所需的软硬件设备如下表。

表1考点提供的主要设备及软件表

序号	设备、软件名称	规格/技术参数、用途	备注
1	大数据技术实训机房	测试场地	保证参考人员有足够间距
2	计算机	CPU酷睿i5以上，内存4G以上，win7/win10操作系统	用于软件开发和软件部署，每人一台
3	Office		编写文档
4	SQLServer2008或以上、Oracle10g或以上、MySQL5.5或以上	数据库管理系统	参考人员任选一种数据库管理系统

附录3评价标准

表2考核评价细则表

评价项	分值	评分细则
数据库创建	10分	没有成功创建数据库，扣5-8分。
数据表创建	25分	数据表创建不成功每一项扣3-5分，字段创建不符合要求每一项扣2-3分，扣完为止。
约束及关系创建	15分	约束创建不成功每一项扣3-5分，关系创建不符合要求每一项扣5分，扣完为止。
数据访问	25分	没有正确写出SQL语句每一项扣4-5分，扣完为止。

数据库管理系统配置与使用		5分	数据库服务器与管理工具配置不正确，无法连接数据库扣5分。
文档规范	数据库名命名规范	2分	数据库名命名不规范扣2分。
	数据表名命名规范	3分	数据表命名不规范每张表扣1分，扣完为止。
	字段名命名规范	5分	字段名命名不规范每项扣0.5分，扣完为止。

表3职业素质评分细则表

序号	评分项	分值	评分 细则
1	代码书写格式规范	3分	代码缩进不规范扣1分、方法划分不规范扣1分、语句结构不规范扣1分（如一行编写两个语句）、使用空行不规范扣1分，扣完为止。
2	注释规范	2分	整个项目没有注释扣2分、有注释，但注释不规范扣1分，扣完为止。
5	运行正确	5分	所写代码无法正常运行扣5分。

二、岗位核心技能模块

模块一 hadoop平台与组件

1.试题编号：3-1 服务器基础网络环境搭建模块

（1）任务描述

某IT科技公司要某市新上线大数据项目，决定在某市甲方公司搭建Hadoop大数据平台，现决定搭建服务器基础网络环境，满足现有大数据规模，同时满足后期业务需求扩展升级。请完成以下任务实验操作。

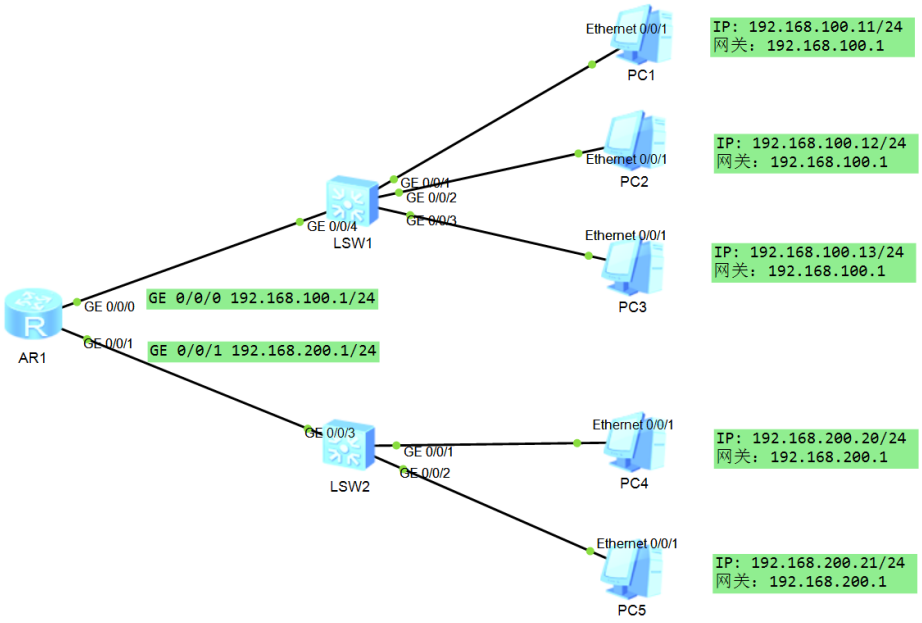


图3.1.1网络拓扑图

任务一：准备计算机和网络设备。（20 分）

将 5 台计算机和2台交换机1台路由器）连接起来。你需要用网线将计算机分别连接到交换机的不同端口上，以构成一个局域网。

任务二：配置路由器。（40 分）

要求：

打开路由器的管理界面，并进行基本的网络配置，包括GE 0/0/0 IP 地址：192.168.100.1、子网掩码：255.255.255.0,GE 0/0/0 IP 地址：192.168.200.1、子网掩码：255.255.255.0

任务三：配置计算机，测试网络连通。（20分）

要求: 设置 PC 端 IP 地址和网关，DNS 统一为：114.114.114.114， 8.8.8.8

（2）作品提交要求见本模块附录1

（3）实施条件要求见本模块附录2

（4）评价标准见本模块附录3

2.试题编号：3-2 配置和管理服务器和存储模块

（1）任务描述

公司新采购了一批 2U 2 路机架服务器(型号 H22H-05)，为某项目做准备。现要对服务器进行安装、配置和管理物理服务器和存储设备，以支持各种企业应用程序。

任务一：升级服务器内存和硬盘，并测试新硬件是否正常工作（10 分）。

要求：根据服务器型号，把服务顺内存升级到128G,硬盘升为至4T以上，查询服务器资料文档，选择内存型号和条数，硬盘型号和个数。

任务二：对服务器的多个硬盘(至少 5 个硬盘)，创建 RAID6,管理磁盘阵列，以提高数据安全性和性能（20 分）。

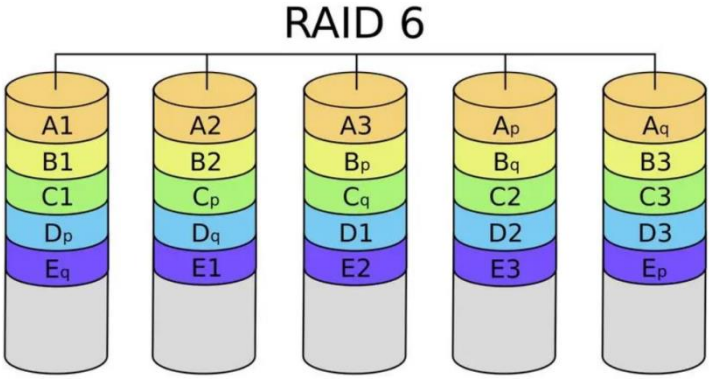


图 3.2.1 磁盘 RAID6

任务三：对现有阵列磁盘进行分区，并使 LVM 对磁盘分区进行管理，用于安装 Centos7 操作系统。（20 分）

要求：

①.Parted 工具对磁盘进行分区，第一个分区为 500M,用于挂载/boot 分区

②.剩下磁盘空间划为一个分区，并使用 LVM 对该分区进行管理

③.根据 Cetonos7 的最小系统要求，使用 LVM 逻辑分区/:20G、swap:8G、/home:20G、/opt:1T，剩余空间不划分作为预留，用于 LVM 扩容需求。

任务四：将物理服务器安装 Centos7 操作系统，并完成服务基础配置（30 分）；

要求:

①.使用 CentOS-7-x86_64-Minimal-1810.iso 镜像文件最小化安装 Centos7 操作系统。分区规划使用任务三。

②.完成 Cetnos7 操作系统安全设置，关闭不必要的服务和端口、启用防火墙、安装和更新安全补丁和软件包、配置 SELinux

③.完成 Cetnos7 操作系统用户账户设置，创建 hadoop 用户，设置密码为：hadoop@zuxia.com

④.完成 Cetnos7 操作系统网络设置：配置网络接口 IP:192.168.100.11/24、网关：192.168.100.1、DNS1:114.114.114.114、DNS2:8.8.8.8 ，主机名:hadoop1

⑤.完成 Cetnos7 操作系统 SSH 设置:配置 SSH 访问服务。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

3.试题编号：3-3 Hadoop 平台安装搭建模块

（1）任务描述

某互联网公司打算搭建自己的大数据平台，用于项目业务大数据分析与开发。主要使用 Apache Hadoop 生态圈技术。请按以下任务要求完成并实现。

节点	IP	角色
master	192.168.100.11	NameNode,DataNode,ResourceManger,NodeManger
slave1	192.168.100.12	DataNode,NodeManger

salve2	192.168.100.13	DataNode,NodeManger

任务一：Hadoop 平台服务器基本环境准备(20 分)

要求：

①.完成三台服务器安装 Cetnos7 操作系统（可以使用虚拟实现）

②.配置三台服务器主机名 IP: master:192.168.100.11,slave1:192.168.100.12,slave2:192.168.100.13，并完成 hosts 映射。

③.创建 hadoop 用户，三台互设置免密登陆。

任务二：JDK 安装与配置。(20 分)

要求：

①.下载 JDK1.8 的 tar 包。

②.把 JDK 安装到/opt/jdk-1.8.xxx 目录下

③.在/etc/profile 中配置 JDK 环境变量，并生效验证

任务三：Hadoop 分布式集群安装与配置（40 分）

要求：

①.下载 Hadoop3.1.3 安装包,并上传至 master 节点

②.安装并配置 Hadoop 系统环境

③.配置 Hadoop 配置文件

④.初始化 HDFS 分布式文件系统。

⑤.启动 hadoop 集群，查看 jps 进程验证集群

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

4.试题编号：3-4 HDFS 模块

（1）任务描述

公司需要设计一个可以容纳大量图像、视频和音频文件的 HDFS 文件系统,用于存储 100T 的数据。请完成以下任务。

任务一：HDFS 设计与优化（10 分）

要求：

①.搭建一个包括 3 个名称节点和 5 个数据节点的 HDFS 集群

②.优化 HDFS 块大小，根据数据访问模式，调整块大小以提高读写性能

任务二：手动上传数据至 HDFS 分布式文件系统(20 分)

要求：

①.在 HDFS 上创建存储数据目录，/data/image/, /data/video/, /data/audio/

②.安将/data/image 目录文件副本数设置为 3，/data/video 目录文件副本数设置为 2。

③.分别把图像、视频和音频文件上传至 HDFS 对应的目录中

任务三：编写 Python 程序实现对 HDFS 分布式文件系统中读操作。(25 分)

要求：

①.pip 安装 HDFS 模块

②.编写程序查看 HDFS 文件系统中的/data/image/下的所有文件列表。

③.编写程序读取 HDFS 文件系统中的/data/image/某一个文件，转存导出到本地文件系统中。

任务四：编写 Python 程序实现对 HDFS 分布式文件系统中写操作。(25 分)

要求：

①.pip 安装 HDFS 模块

②.编写 Python 程序查看 HDFS 文件系统中的/data/video/下的所有文件列表。

③.编写 Python 程序，把本地一个视频上传至 HDFS 文件系统中的/data/video 中，要求上传的视频文件副本数为 2。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

5.试题编号：3-5 分布式数据库模块

（1）任务描述

某互联网公司打算使用分布式数据库，用于存储电商数据，对电商数据进行数据分析，请完成以任务。

任务一：HBase 安装与配置(20 分)

要求：

①.下载 HBase2.2.3 软件 tar 包

②.解压安装至/opt/hbase2.23/目录下。

③.修改 HBase 配置文件，zookeeper 使用 HBase 自带的。

任务二：根据以表分布式电商数据库设计(20 分)

商品表（Products）

列族	列名	数据类型	描述
info	product_id	string	商品 ID，唯一标识符
info	product_name	string	商品名称
info	product_brand	string	商品品牌
info	product_price	double	商品价格
info	product_weight	double	商品重量
detail	product_detail	binary	商品详情信息，包括图文和 HTML 格式
image	product_image_1	binary	商品图片 1，二进制格式
image	product_image_2	binary	商品图片 2，二进制格式
image	product_image_3	binary	商品图片 3，二进制格式
review	review_id	string	商品评价 ID，唯一标识符
review	review_title	string	商品评价标题
review	review_content	string	商品评价内容
review	review_score	int	商品评价得分，0-5 分

订单表（Orders）

列族	列名	数据类型	描述
order	order_id	string	订单 ID，唯一标识符
order	user_id	string	用户 ID
order	create_time	long	订单创建时间，Unix 时间戳
product	product_id	string	商品 ID
product	product_count	int	商品数量
product	product_price	double	商品单价
detail	delivery_name	string	收货人姓名
detail	delivery_phone	string	收货人电话号码
detail	delivery_addr	string	收货地址

用户表（Users）

列族	列名	数据类型	描述
info	user_id	string	用户 ID，唯一标识符
info	user_name	string	用户昵称
info	user_email	string	用户邮箱
info	user_phone	string	用户电话号码
info	user_address	string	默认收货地址
detail	user_password	string	用户密码
order	order_id	string	订单 ID，按时间倒序排

要求：

①.创建 ecommerce 命名空间

②.在 ecommerce 命名空间中创建用户表，定单表，商品表

任务三：在 HBase 电商数据仓中添加数据(20 分)

商器表数据：

```
product_id  product_name  product_brand product_price product_weight  product_detail
    product_image_1  product_image_2  product_image_3  review_id  review_title
    review_content  review_score
P001  iPhone 12  Apple  7999.00  164 g  Binary Binary Binary Binary R001  非 常 好 用
    这款手机非常好用，操作流畅，摄像效果也很棒！  5
P002  Galaxy S21  Samsung  6999.00  171 g  Binary Binary Binary Binary R002  太贵了
```

	这款手机性能还不错，但价格实在是太高了。 3									
P003	ThinkPad X1 Carbon	Lenovo	9999.00	1.13 kg	Binary	Binary	Binary	Binary	R003	
	轻薄便携	这台笔记本轻薄便携，适合商务人士出差使用。	4							
P004	AirPods Pro	Apple	1999.00	45.6 g	Binary	Binary	Binary	Binary	R004	降噪效果好
	这款耳机的降噪效果非常好，可以过滤掉大部分外界噪音。 5									
P005	PlayStation 5	Sony	4499.00	4.5 kg	Binary	Binary	Binary	Binary	R005	很流畅 这个游戏机的运行速度很流畅，游戏效果也非常棒！ 5

订单表:

order_id	user_id	create_time	product_id	product_count	product_price	delivery_name	delivery_phone	delivery_addr
0001	U001	1625101200000	P001	1	7999.00	张三	133xxxx5678	北京市朝阳区 xxx 路 xx 号
0002	U002	1625187600000	P003	1	9999.00	李四	139xxxx1234	上海市 xxx 路 xx 号
0003	U003	1625274000000	P002	1	6999.00	王五	158xxxx8765	广州市 xxx 路 xx 号
0004	U004	1625360400000	P001	2	7999.00	赵六	186xxxx4321	深圳市 xxx 路 xx 号
0005	U005	1625446800000	P004	1	1999.00	刘七	177xxxx9876	成都市 xxx 路 xx 号
0006	U006	1625533200000	P005	1	4499.00	钱八	136xxxx3456	武汉市 xxx 路 xx 号
0007	U007	1625619600000	P001	1	7999.00	孙九	158xxxx5432	上海市 xxx 路 xx 号
0008	U008	1625706000000	P002	1	6999.00	李十	177xxxx1234	北京市海淀区 xxx 路 xx 号
0009	U009	1625792400000	P003	1	9999.00	刘强东	186xxxx8765	北京市朝阳区 xxx 路 xx 号
0010	U010	1625878800000	P005	2	4499.00	马化腾	133xxxx0987	深圳市南山区 xxx 路 xx 号
0011	U001	1625965200000	P004	1	1999.00	董明珠	139xxxx5678	广州市 xxx 路 xx 号
0012	U003	1626051600000	P001	1	7999.00	罗永浩	150xxxx2345	深圳市福田区 xxx 路 xx 号
0013	U005	1626138000000	P002	2	6999.00	杨幂	186xxxx8765	北京市朝阳区 xxx 路 xx 号
0014	U007	1626224400000	P005	1	4499.00	范冰冰	133xxxx3456	上海市 xxx 路 xx 号
0015	U009	1626310800000	P003	1	9999.00	刘备	177xxxx2345	成都市 xxx 路 xx

用户表:

user_id	user_name	user_email	user_phone	user_address	user_password	order_id
001	John Smith	john.smith@example.com	+1 555-123-4567	123 Main St, Apt 4A	password123	ORD00001
002	Jane Doe	jane.doe@example.com	+1 555-234-5678	456 Oak Ave	abc123	ORD00002
003	Bob Johnson	bobjohnson@example.com	+1 555-345-6789	789 Elm St	mypassword	ORD00003
004	Lisa Brown	lisabrown@example.com	+1 555-456-7890	321 Maple Dr	p@ssw0rd	ORD00004, ORD00005

005	David Lee	david.lee@example.com	+1 555-567-8901	654 Pine Rd	password123
006	Sarah Kim	sarah.kim@example.com	+1 555-678-9012	987 Cedar Ln	qwerty123
007	Tom Wilson	tomwilson@example.com	+1 555-789-0123	246 Cherry St	letmein123 ORD00006, ORD00007
008	Emily Chen	emilychen@example.com	+1 555-890-1234	135 Oak Ave	password123
009	Kevin Wang	kevin.wang@example.com	+1 555-901-2345	864 Elm St, Apt 2B	soccer123 ORD00008
010	Grace Wu	grace.wu@example.com	+1 555-012-3456	753 Maple Dr	iloveyou123 ORD00009, ORD00010

要求：

①.在 HBase shell 中分别给商品表，定单表，用户表各添加以上模拟数据

②.自行构造设计合理的 RowKey

任务四：编写 python 程序，读取三个 csv 文件，添加到 HBase 数据对应的表中（20 分）

要求：

①.编写 python 程序完成读取 CSV 文件数据,文件内容为任务三中的三个表数据。

②.使用 happybase - Python HBase 客户端库。

③.自行构造设计合理的 RowKey，完成对商品表，定单表，用户表的添加到数据库中。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

6.试题编号：3-6 Flume 日志采集模块

（1）任务描述

公司要对某个应用项目上的服务器，做日志收集处理。在 WEB 应用服务器上有一个日志文件/var/log/myapp.log，需要将服务器日志内容定期收集并存储到 HDFS 中。

任务一：安装配置分布式日志采集工具 Flume。（30 分）

要求：

①.安装版本：apache-flume-1.9.0

②.解压安装在/opt/目录下

③.在/etc/profile 中配置 Flume 环境变量

任务二：编写 flume 脚本配置，完成采数据至 HDFS。（50 分）

要求：

①.采集服务日志文件/var/log/myapp.log，存入至 HDFS 中

②.存放 HDFS 中的/data/logs 目录

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

7.试题编号：3-7 Maxwell 日志采集模块

（1）任务描述

公司要对某个应用项目上的服务器，做日志收集处理。在 MySQL 服务器数据库的二进制日志，需要将服务器日志内容定期收集并存储到 Kafka 中。

任务一：安装配置 Maxwell 工具。（30 分）

要求：

①.安装版本：maxwell-1.29.2

②.解压安装在/opt/目录下

③.在/etc/profile 中配置 Flume 环境变量

任务二：编写 maxwell 脚本配置，完成采数据至 Kafka。（50 分）

要求：

①.采集 MySQL 服务二进制日志文件，存入至 Kafka 中 topiclog 主题中。

②.存放 Kafka 集群:master:9092,slave1:9092,slave2:9092(Kafka 集群根据实际情况自行调整)。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

8.试题编号：3-8 数据仓库工具模块

（1）任务描述

某大数据公司为方便计算分析电商数据，采用 Hive 数据仓库技术，来建模设计电商数据仓库,实现管理和分析电子商务业务的数据。请完成以下任务。

任务一：安装配置 HIVE。（20 分）

要求：

①.下载 hive-3.1.2 软件 tar 包

②.解压安装至/opt/hive-3.1.2/目录下。

③.Hive 元数据使用 MySQL5.7 进行存储，MySQL 账号:hive，密码为：123456

④.在/etc/profile 中配置 Hive 环境变量

任务二：电商数据仓库设计（20 分）

要求：根据以下表及表结构，采用星模型来设计数仓模型。

订单数据表(Order Table)：

字段	数据类型	说明
order_id	INT	订单 ID
customer_id	INT	客户 ID
product_id	INT	商品 ID
order_date	DATE	下单日期

字段	数据类型	说明
order_total_amount	DECIMAL(10,2)	订单总金额
order_status	VARCHAR(10)	订单状态（已支付/已取消）

销售数据表(Sales Table)：

字段	数据类型	说明
product_id	INT	商品 ID
sales_date	DATE	销售日期
sales_quantity	INT	销售数量
sales_amount	DECIMAL(10,2)	销售金额
sales_customer_count	INT	销售客户数量
sales_region	VARCHAR(10)	销售区域

客户数据表(Customer Table)：

字段	数据类型	说明
customer_id	INT	客户 ID
customer_name	VARCHAR(20)	客户名称
customer_gender	VARCHAR(4)	客户性别
customer_age	INT	客户年龄
customer_region	VARCHAR(10)	客户所在区域
customer_email	VARCHAR(50)	客户电子邮箱

支付数据表(Payment Table)：

字段	数据类型	说明
order_id	INT	订单 ID
payment_date	DATE	支付日期

字段	数据类型	说明
payment_amount	DECIMAL(10,2)	支付金额
payment_method	VARCHAR(20)	支付方式
payment_status	VARCHAR(10)	支付状态（已支付/未支付）

任务三：电商数仓分层设计（20 分）

要求：根据电商数据设计会层：ODS 层,DWD 层,DWS 层,DWT 层,DIM 层,ADS 层。

①.在 Hive 中创建对应的层数据库。

②.在各层中设计对应的表及表结构，并在 Hive 中创建表。

任务四：数仓加截数据。（20 分）

要求：把已经存在的电商数据加载到 Hive 数据仓中 ODS 层中。

①.数据存在本地/data/ec/目录下和 HDFS 中的/data/etc/目录下

②.数据文件中各字段是以“，”分隔

③.使用 load data 命令加载数据

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

9.试题编号：3-9 HQL 语句模块

（1）任务描述

电商平台，数据表结构如下，请根据任务要求完成以下任务。

Product 表

列名	数据类型	说明
id	INT	商品 ID
name	VARCHAR(255)	商品名称
price	DECIMAL(10,2)	商品单价
stock	INT	商品库存量

Order 表

列名	数据类型	说明
id	INT	订单 ID
userId	INT	下单用户的 ID
totalPrice	DECIMAL(10,2)	订单总金额
createTime	TIMESTAMP	订单创建时间

User 表

列名	数据类型	说明
id	INT	用户 ID
name	VARCHAR(255)	用户姓名
email	VARCHAR(255)	用户邮箱地址

OrderItem 表（记录每个订单中包含的商品信息）

列名	数据类型	说明
orderId	INT	订单 ID
productId	INT	商品 ID
quantity	INT	商品数量
price	DECIMAL(10,2)	商品单价
totalPrice	DECIMAL(10,2)	商品总价

任务一：创建数据库和表（20 分）

要求：

①.在 Hive 中创建 ec 数据库

②.分别外部表来创建 Product（商品）、Order（订单）和 User（用户）表

任务二：对三个外部表加载数据（20 分）

要求：

①. 加载数据自行模拟,模拟数据合理，使用 python 脚本编程生成。

②.数据大小不能少于 10M。

任务三：对电商数据表进行 HQL 操作完成以下要求。（40 分）

要求：

①.按照商品价格从高到低排序

②.统计每个用户的订单总金额

③.扣减某个商品的库存

④.查询某个用户最近购买的商品列表

⑤.删除所有库存为 0 的商品

⑥.查询每个商品的销售量和销售额

⑦.统计每个用户最近 7 天的订单总金额

⑧.查询所有购买了某个商品的用户信息

⑨.统计每种商品的平均价格和库存量

⑩.删除所有没有订单的商品

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

10.试题编号：3-10 PySpark 模块

（1）任务描述

公司原有的 Hive 离线分析数据，特别是处理数据大时，查询慢，为解决这问题，公司换 Spark on hive 作为计算引擎。通过在 Spark 上执行 Hive SQL 查询来提高查询性能。相对于传统的 MapReduce 引擎，Spark 具有更好的性能和效率。请完成以下任务:

任务一：PySpark 安装与搭建。（30 分）

要求：

①.安装版本:Python3.8,Spark3.1.1

②.Spark 以 Standalone 集群模式安装

③.安装目录为:/opt/目录下

④.在/etc/profile 中配置 Spark 环境变量

任务二： PySpark 程序抽取数据（50 分）

要求：

①.编写 PySpark 程序来实现

②.全量抽收 MySQL 中的数据至 Hive 的 ods 层表中

③.增量数据抽(T+1)取 MySQL 中的数据至 Hive 的 ods 层表中

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

Hadoop 平台与组件模块附录

附录1作品提交

答案以“答题文件”的形式提交。请按以下要求创建答题文件夹和答题文件：

①创建答题文件夹

创建以”考生号_题号”命名的文件夹，存放所有答题文件，例如：“340103**********_3_1\”。

②创建答题文件

a.项目源文件

创建任务子文件夹，如：“340103**********_3_1\task_1\”，存放任务一所有结果。

b.页面截图文件

在任务子文件夹中，存放截图.doc文件，它用于保存安装，配置，启动，运行执行过程中的屏幕截图，每张截图中每个关键配置或结果等，必须用红色矩形框标识出来并加以文字说明。

③提交答题文件

将”考生号_题号”文件夹打包，形成“考生号_题号.RAR”文件，如：“340103**********_3_1.rar”，将该文件按要求进行上传。

④考核时量

考核时间为180分钟。

附录2实施条件

所需的软硬件设备如下表。

表1考点提供的主要设备及软件表

序号	场地、设备、软件名称	规格/技术参数、用途	备注
1	大数据技术实训机房	测试场地	保证参考人员有足够间距
2	服务器	CPU酷睿i5以上，内存16G以上，CeonOS7操作系统,Docker-ce23.1	可以高性能PC机代替，用于安装服务器基础组件，每人一台
3	计算机	CPU酷睿i5以上，内存4G以上，win7/win10/Ubuntu操作系统	用于软件开发和软件部署，每人一台
4	Office、WPS		编写文档

附录3评价标准

评分项目一：实操文档（10分）

表2 实操文档评分细则表

序号	评分项	分值	评分细则
1	实操文档有无	2分	有实操文档得分，无实操文档扣2分。

2	文档任务截图	4分	有操作过程截图得分，无操作过程截图扣4分。
3	文档任务截图标注	4分	有文档任务截图标注说明和画框得分，无标注和画框扣4分

评分项二：依据题的任务，完成任务功能（80分）

表3 项目功能评分细则表

序号	评分项	分值	评分细则
1	功能实现	80分	试题按任务分值评分;根据各任务功能,完成并实现，截图体现关键信息判断，符合要求得该分值，否则不得分。子任务有具体要求,若未完成子任务要求则扣对应子任务3-5分。

评分项三：职业素质（10分）

表4 职业素质评分细则表

序号	评分项	分值	评分细则
1	代码书写格式规范	3分	代码缩进不规范扣1分、方法划分不规范扣1分、语句结构不规范扣1分（如一行编写两个语句）、使用空行不规范扣1分，扣完为止。
2	注释规范	2分	整个项目没有注释扣2分、有注释，但注释不规范扣1分，扣完为止。
3	端口配置	1分	端口号配置不正确，扣1分。
4	部署正确	4分	项目代码未正确配置到指定服务器目录下，扣4分。

模块二 数据处理技术

1.试题编号：4-1：淘宝数据采集模块

（1）任务描述

你所在的电商公司最近打算进行市场竞争分析，需要对各大电商平台上的商品价格和评价数据进行统计分析。因此，你需要编写一个程序来采集淘宝上商品的价格和评论数据

任务一：采集淘宝评论数据

假设你是一家电商公司的数据分析师，你想要收集指定商品在淘宝上的价格和评论数据。请编写一个Python 程序，使用 selenium 和 BeautifulSoup 库通过模拟浏览器行为获取目标数据，并将结果保存到 CSV 文件中

要求：

①.目标商品为“Apple AirPods Pro”。(10 分)

②.获取商品的价格、评分（1 分非常不满意；；一般 3 分； 5 分非常满意）、评论内容、评论时间。(30 分)

③.仅获取前 10 页的评论数据。(20 分)

④.结果保存到 CSV 文件中，文件名为“Apple_AirPods_Pro.csv”，文件编码为 UTF-8。(20 分)

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

2.试题编号：4-2：京东数据采集模块

（1）任务描述

你所在的电商公司最近打算进行市场竞争分析，需要对各大电商平台上的商品价格和评价数据进行统计分析。因此，你需要编写一个程序来采集京东上商品的价格和评论数据。

假设你是一家电商公司的数据分析师，你想要收集指定商品在淘宝上的价格和评论数据。请编写一个Python 程序，使用 selenium 和 BeautifulSoup 库通过模拟浏览器行为获取目标数据，并将结果保存到 CSV 文件中

任务一：采集京东评论数据

假设你是一家电商公司的数据分析师，你想要收集京东平台上所有 iPhone 12 手机的价格和评论数据。请编写一个 Python 程序，使用 requests 和 BeautifulSoup 库获取目标数据，并将结果保存到 Excel 文件中

要求：

①.目标商品为“iPhone 12”。 (10 分)

②.获取商品的价格、评论数、评分(1-5 星对应 1-5 分)、评论内容、评论时间。(30 分)

③.仅获取前 100 个商品的评论数据。(20 分)

④.结果保存到 Excel 文件中，文件名为“iPhone_12.xlsx”，文件编码为 UTF-8。(20 分)

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

3.试题编号：4-3：电商订单数据 ETL 模块

（1）任务描述

某电商平台需要对其订单数据进行 ETL，以方便后续的数据分析工作，你作为数据工程师需要完成以下任务。

任务一：电商订单数据清洗与提取

某电商平台提供的订单数据，需要进行清洗和提取，将满足以下条件的订单信息筛选出来：

- 订单状态为“完成”，“关闭”或“退款成功”；
- 订单金额大于 500 元；
- 下单时间在 2022 年 1 月 1 日后。

订单数据集包含以下字段：

字段名称	字段类型	字段描述
order_id	string	订单编号，唯一标识一笔订单
user_id	string	用户编号，唯一标识一个用户
status	string	订单状态，取值范围为“待付款”，“已付款”，“待发货”，“已发货”，“已签收”，“退款中”，“退款成功”，“关闭”，“完成”
amount	float	订单金额，单位为元
create_time	timestamp	订单创建时间，格式为 yyyy-MM-dd HH:mm:ss

数据集示例：

order_id	user_id	status	amount	create_time
order001	user001	待付款	568.50	2022-01-01 10:00:00
order002	user002	已发货	236.80	2022-02-01 14:30:00
order003	user003	关闭	99.90	2022-03-01 09:45:00
order004	user004	退款成功	1200.0	2022-04-01 19:20:00

要求：

①.使用 ETL 工具或编程语言，读取并清洗原始数据集；(30 分)

②.按照题目描述的条件，筛选符合条件的订单记录；(30 分)

③.将结果导出至 CSV 文件中，包含订单编号、用户编号、订单状态、订单金额、订单创建时间等字段。(20 分)

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

4.试题编号：4-4：电商商品数据 ETL 模块

（1）任务描述

某电商平台需要对其商品数据,进行 ETL，以方便后续的数据分析工作，你作为数据工程师需要完成以下任务。

任务一：电商商品数据清洗与转换

某电商平台提供的商品数据，需要进行清洗和转换，将商品信息转换为标准的商品属性，包括“品牌”、“型号”、“颜色”、“尺寸”等属性，以方便后续的数据分析工作。

商品数据集包含以下字段：

字段名称	字段类型	字段描述
item_id	string	商品编号，唯一标识一个商品
item_name	string	商品名称
item_desc	string	商品描述

数据集示例：

item_id	item_name	item_desc
item001	美的空调	美的（Midea) 手持遥控 空调挂机...
item002	华为手机	华为（HUAWEI) Mate 40 5G 手机...
item003	清扬洗发水	清扬（CLEAR)男士去屑洗发露...

要求：

①.使用 ETL 工具或编程语言，读取并清洗原始数据集；(30 分)

②.将商品名称字段拆分为品牌、型号、颜色、尺寸等属性，需自行确定拆分规则；(30 分)

③.将结果导出至 CSV 文件中，包含商品编号、品牌、型号、颜色、尺寸等字段；(20 分)

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

5.试题编号：4-5：员工满意度数据清洗模块

（1）任务描述

某公司通过调查员工满意度，得到了一个 Excel 文件，该文件包含有关公司员工的基本信息、满意度等数据，但是这些数据存在许多错误和缺失值。现在需要对这些数据进行清洗。

Excel 文件：employee_data.xlsx

员工编号	姓名	年龄	性别	邮箱	手机号码	国家	城市	地址	满意度
E0001	李明	25	男	LI.MING@EXAMPLE.COM	+861234567890	中国	北京	朝阳区	80
E0002	张三	30	男	zhangsan@example.com	13123456789	中国	上海	浦东新区	85
E0003	王五	-2	男	wangwu@example.com	021-12345678	中国	广州	越秀区	75
E0002	张三	30	男	zhangsan@example.com	13123456789	中国	上海	浦东新区	85
E0004	李四	35		Lisi@example.com	+852-12345678	香港	中西区	德辅道中	
E0005	小兰	25	女	xiaolan@example.com	0755-23456789	中国	深圳	罗湖区	86
E0006	Lucy		女	LUCY@EXAMPLE.COM	020-34567890	美国	纽约	曼哈顿	78

任务一：删除重复行(10 分)

要求：删除 employee_data.xlsx 重复行记录数据,保留一行。

任务二：填充缺失值(20 分)

要求：

①.如果数据集中缺失值的比例较小，可以考虑直接删除这些缺失值所在的行或列；

②.年龄，采用填充平均值或者中位数的方式来填充缺失值；

③.性别，可采用众数来进行填充。

任务三：删除非法值(20 分)

要求：

①.如年龄小于 0 或大于 100 的记录,使用平均值或者中位数填充。

②.年龄，采用填充平均值或者中位数的方式来填充缺失值。

任务 4：数据转换(10 分)

要求：将电子邮件地址转换为小写格式。

任务 5：格式化电话号码(20 分)

要求：（如 +86(10)12345678 转换为 010-12345678 的格式）。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

6.试题编号：4-6：患者病历记录数据清洗模块

（1）任务描述

某医院收集了一份患者病历记录，其中包含了很多错误、缺失和异常值。现在需要对这些数据进行清洗。使用 Python 完成以下数据清洗任务，并将结果保存到新的 CSV 文件中。

数据集 CSV 文件：patient_records.csv

序号	病人姓名	生日	性别	联系方式	地址	医生编号	检查时间	费用
1	张三	19891201	male	13800000001	北京市海淀区	001	20220101102030	$60.0
2	李四	19951002	female	13900000002	河北省石家庄市	002	20220303153000	40.0
3	王五	20210230	male	13100000003	北京市朝阳区	003	20220404133000	80.0
4	赵六	19871215	female	15800000004	广东省深圳市福田区	004	20220408193030	$125.0
5	小七	19990101		13600000005	上海市黄浦区	005	20220505123045	70
6	皮八	19931002	unknown		湖南省长沙市	006	20220606103030	$100.0
7	王九		male	15200000007	江苏省南京市		20220707153000	$110.0
8	钱十	19950909	female	+8618000000010	上海市浦东新区	008	20220808123045-0800	90.0

任务一：删除非法值(10 分)

要求：删除生日列中不合法的日期（如 20210230）

任务二：填充缺失值(20 分)

任务三：格式化日期时间列(10 分)

要求：格式化检查时间列，将其转换为 yyyy-mm-dd HH:MM:SS 的格式；

任务 4：将性别列转换为数字编码，并填充缺失值(10 分)

要求：将性别列中的 male 转换为 0，female 转换为 1，并填充缺失值；

任务 5：将医生编号列转换为字符串格式(10 分)

要求：将医生编号列中的数字前补零（如 001、002），并填充缺失值；

任务 6：删除重复行(10 分)

任务 7：将费用列转换为浮点数类型。(10 分)

要求：

①. 统计货币单位；

②. 费用列转换为浮点数类型；

（2）作品提交要求见本模块附录 1

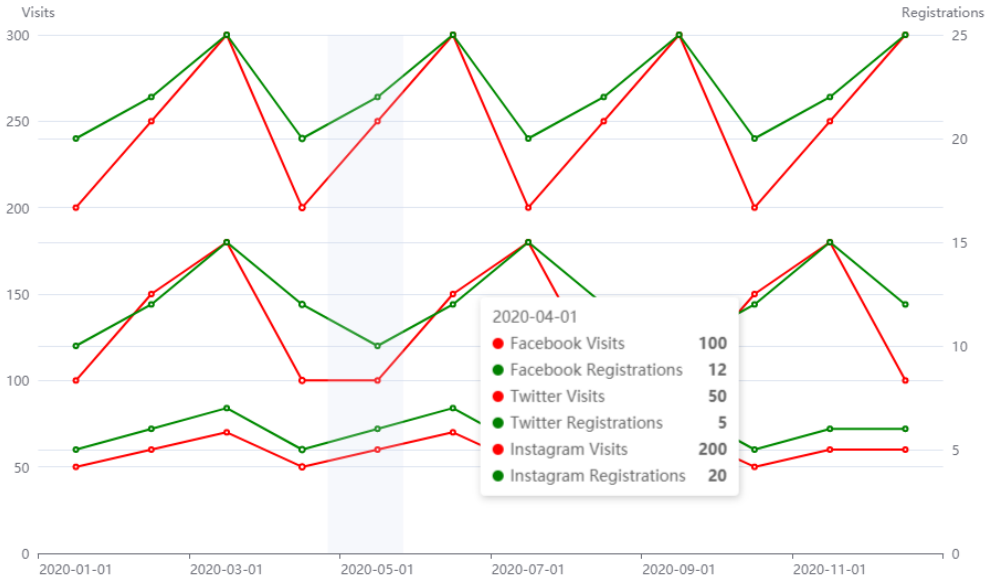
（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

7. 试题编号：4-7：数据可视化折线图模块

（1）任务描述

你正在帮助一家社交媒体公司分析他们的用户数据，他们想要了解用户在不同平台上的行为。创建一个可交互的多轴折线图，其中每条线表示一个平台上的用户行为，如访问次数、注册用户数量等，并使用不同颜色为每个平台的每条线条分配不同颜色。使用 Echarts 或 Highcharts 实现可视化展示。



数据集：下面是一个示例数据集，包含了一些假想的社交媒体平台和它们的数据：

```
{
        "Facebook": [
          { "date": "2020-01-01", "visits": 100, "registrations": 10 },
          { "date": "2020-02-01", "visits": 150, "registrations": 12 },
          { "date": "2020-03-01", "visits": 180, "registrations": 15 },
          { "date": "2020-04-01", "visits": 100, "registrations": 12 },
          { "date": "2020-05-01", "visits": 100, "registrations": 10 },
          { "date": "2020-06-01", "visits": 150, "registrations": 12 },
          { "date": "2020-07-01", "visits": 180, "registrations": 15 },
          { "date": "2020-08-01", "visits": 100, "registrations": 12 },
          { "date": "2020-09-01", "visits": 100, "registrations": 10 },
          { "date": "2020-10-01", "visits": 150, "registrations": 12 },
          { "date": "2020-11-01", "visits": 180, "registrations": 15 },
          { "date": "2020-12-01", "visits": 100, "registrations": 12 }
        ],
```

```
        "Twitter": [
          { "date": "2020-01-01", "visits": 50, "registrations": 5 },
          { "date": "2020-02-01", "visits": 60, "registrations": 6 },
          { "date": "2020-03-01", "visits": 70, "registrations": 7 },
          { "date": "2020-04-01", "visits": 50, "registrations": 5 },
          { "date": "2020-05-01", "visits": 60, "registrations": 6 },
          { "date": "2020-06-01", "visits": 70, "registrations": 7 },
          { "date": "2020-07-01", "visits": 50, "registrations": 5 },
          { "date": "2020-08-01", "visits": 60, "registrations": 6 },
          { "date": "2020-09-01", "visits": 70, "registrations": 7 },
          { "date": "2020-10-01", "visits": 50, "registrations": 5 },
          { "date": "2020-11-01", "visits": 60, "registrations": 6 },
          { "date": "2020-12-01", "visits": 60, "registrations": 6 }
        ],
        "Instagram": [
          { "date": "2020-01-01", "visits": 200, "registrations": 20 },
          { "date": "2020-02-01", "visits": 250, "registrations": 22 },
          { "date": "2020-03-01", "visits": 300, "registrations": 25 },
          { "date": "2020-04-01", "visits": 200, "registrations": 20 },
          { "date": "2020-05-01", "visits": 250, "registrations": 22 },
          { "date": "2020-06-01", "visits": 300, "registrations": 25 },
          { "date": "2020-07-01", "visits": 200, "registrations": 20 },
          { "date": "2020-08-01", "visits": 250, "registrations": 22 },
          { "date": "2020-09-01", "visits": 300, "registrations": 25 },
          { "date": "2020-10-01", "visits": 200, "registrations": 20 },
          { "date": "2020-11-01", "visits": 250, "registrations": 22 },
          { "date": "2020-12-01", "visits": 300, "registrations": 25 }
        ]
}
```

任务一：加载数据集(30 分)

要求：

①数据格式为 json 格式；

②使用异步加载数据；

任务二：绘制折线图(40 分)

要求：

①创建一个多轴折线图,每一个轴代表一个平台；

②为每一个轴添加相应的数据，例如访问次数或注册用户数量等；

③为每个平台的每条线条分配不同的颜色；

④鼠标悬停在数据点上时，将显示更详细的信息；

任务三：根据所得结果和图表呈现，对用户数据情况进行简要分析和总结

（2）作品提交要求见本模块附录 1

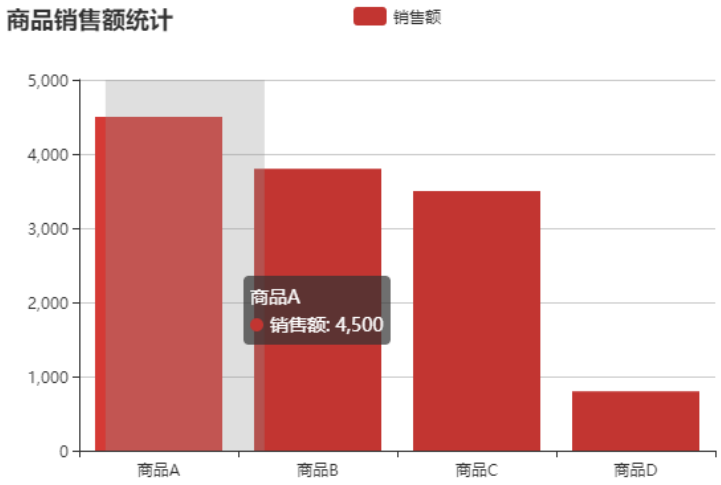
（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

8.试题编号：4-8：数据可视化柱状图模块

（1）任务描述

某电商公司想要对其商品销售情况进行可视化分析，以帮助管理层更好地了解销售情况并制定合理的业务决策。使用 Echarts 实现某段时间内的商品销售额统计并在柱状图中可视化展示。



数据集:数据集可以使用 MySQL 或其他关系型数据库存储。可以使用以下实例数据:

表名：sales

product_id	product_name	sale_time	sale_amount
1001	商品A	2022-01-01 10:00:00	1000
1002	商品B	2022-01-01 10:01:00	2000
1003	商品C	2022-01-01 10:02:00	3000
1001	商品A	2022-01-02 10:00:00	500
1002	商品B	2022-01-02 10:01:00	1000
1003	商品C	2022-01-02 10:02:00	800
1001	商品A	2022-01-03 10:00:00	2000
1002	商品B	2022-01-03 10:01:00	1500
1004	商品D	2022-01-03 10:02:00	800

任务一：分析商品销售数据集。(20 分)

数据集包含以下字段：

商品 ID（product_id）

商品名称（product_name）

商品销售时间（sale_time）

商品销售金额（sale_amount）

任务二：对数据进行处理并计算某段时间内各个商品的销售额总和。(10 分)

任务三：在柱状图中使用销售额作为 Y 轴值，商品名称作为 X 轴类目展示销售情况。(40 分)

要求实现以下功能：

①.统计某段时间内各个商品的销售额。

②.按照销售额从高到低排序。

③.实现可拖拽的数据筛选器，可以按照日期范围或者指定商品进行筛选。(10 分)

任务 4：根据所得结果和图表呈现，对销售情况进行简要分析和总结

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

9.试题编号：4-9：销售预测数据分析模块

（1）任务描述

假设你是某公司的数据分析师，公司想要了解其销售额、利润、市场份额等数据以及预测未来的业绩，并对该数据进行可视化展示和存储。你需要使用数据挖掘工具，完成该任务。

数据集：数据集可以从公司数据库中获取，至少包含时间、销售额、产品类型三个字段。数据量至少有 1000 条记录。也可以使用 Kaggle 平台提供的 Online Retail II Dataset 数据集，其中包含了一个在线零售商店的销售记录，包括订单编号、产品名称、销售额、利润、销售时间和地点等字段。

订单编号	产品名称	销售额	利润	销售时间	地点	产品类别
001	产品 1	1000	200	2020-01-01	北京市	电子产品
002	产品 2	500	50	2020-01-02	上海市	家电
003	产品 3	800	100	2020-01-03	广州市	日用品
…	…	…	…	…	…	…

任务一：收集数据(10 分)

要求：从公司数据库中提取销售历史数据集，至少包含时间、销售额、产品类型三个字段。要求数据集至少有 1000 条记录。

任务二：数据准备(10 分)

要求：使用数据清洗工具处理数据，确保数据集无重复、无缺失值和异常值。

任务三：数据探索(10 分)

要求：使用数据可视化工具展示销售数据的分布、趋势等信息，对销售数据进行探索性分析，并得出探究结果。

任务四：特征工程(10 分)

要求：根据探索性分析结果，选择合适的特征用于建模和预测。

任务五：建模和预测(20 分)

要求：使用数据挖掘工具中的机器学习算法构建销售预测模型，并使用历史数据集进行训练。使用模型预测未来一个月（30 天）的销售情况。

任务六：模型评估(10 分)

要求：评估模型的性能，使用评价指标评估模型的预测误差程度。

任务七：存储数据(10 分)

要求：将清洗好的数据、选择好的特征和建好的模型存储到数据库或文件中，以便后续使用

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

10. 试题编号：4-10：用户评论情感数据分析模块

（1）任务描述

假设您是一家电商企业的数据分析师，您想要了解顾客对该企业的商品的情感反应，以便改善产品和服务质量。为了实现这一目标，您将执行一项情感分析项目，对该企业的社交媒体平台上的用户评论进行情感分析。使用数据挖掘工具对该电商企业的社交媒体用户评论进行情感分析，对不同类型的商品进行分析，发现顾客的最喜爱和最不喜爱的商品。

以下是一个含有中文用户评论数据集的 CSV 文件的示例链接：电商产品中文评论数据集

该数据集包含了约 14 万条中文评论，涉及 10 个不同的商品类别。每个评论都标记有其情感极性（正面、负面或中性），以及所属的商品类别。您可以通过上述链接下载数据集来进行分析和实操

示例数据集：

种类(cat)	标记()label	评论文本（review）
手机	正面(1)	这个产品真的很好用！
手机	负面(0)	这个产品真的很差劲。
书籍	中性(2)	物流有点慢，但产品本身还不错。
手机	正面(1)	这个产品超出了我的期望！
水果	负面(0)	我与客服的经验很不好。
水果	中性(2)	产品损坏了，但客服非常有帮助解决了问题。
…	…	…

任务一：分类(20 分)

要求：

①. 将评论分为正面、中性和负面三个类别。

②. 使用工具：您可以使用一种或多种数据挖掘工具来完成该任务，例如 Python 中的 pandas、scikit-learn 和 NLTK 等

任务二：情感分值 (20 分)

要求：

①. 基于情感分析，给出每条评论的情感分值，评估其评价的程度。

②. 在进行情感分析之前，您需要先对评论数据进行预处理，例如对评论进行清洗、停用词处理和分词等

任务三：商品分析(30 分)

要求：

①. 将评论数据按不同的商品类型进行分组，对每个商品类型进行情感分析，发现顾客最喜爱和最不喜爱的商品；

②. 您可以使用情感词典、机器学习、深度学习等方法来进行情感分析，选择和实现相应的方法

任务 4：分析报告(10 分)

要求：

①.您需要将评论数据按不同的商品类型进行分组，对每个商品类型进行情感分析，并给出产品开发、营销策略等方面的建议；

②.提交一份完整的分析报告，清晰地阐述情感分析过程、结果和结论，提供可视化结果和解释。

（2）作品提交要求见本模块附录 1

（3）实施条件要求见本模块附录 2

（4）评价标准见本模块附录 3

数据处理技术模块附录

附录1作品提交

答案以“答题文件”的形式提交。请按以下要求创建答题文件夹和答题文件：

①创建答题文件夹

创建以”考生号_题号”命名的文件夹，存放所有答题文件，例如：“340103**********_4_1\”。

②创建答题文件

a.项目源文件

创建任务子文件夹，如：“340103**********_4_1\task\”，存放任务所有结果。

b.页面截图文件

在任务子文件夹中，存放截图.doc文件，它用于保存安装，配置，启动，运行执行过程中的屏幕截图，每张截图中每个关键配置或结果等，必须用红色矩形框标识出来并加以文字说明。

③提交答题文件

将”考生号_题号”文件夹打包，形成“考生号_题号.RAR”文件，如：“340103**********_4_1.rar”，将该文件按要求进行上传。

④考核时量

考核时间为180分钟。

附录2实施条件

所需的软硬件设备如下表。

表1 考点提供的主要设备及软件表

序号	场地、设备、软件名称	规格/技术参数、用途	备注
1	大数据技术实训机房	测试场地	保证参考人员有足够间距
2	服务器	CPU酷睿i5以上，内存16G以上，CeonOS7操作系统,Docker-ce23.1	可以高性能PC机代替，用于安装服务器基础组件，每人一台
3	计算机	CPU酷睿i5以上，内存4G以上，win7/win10/Ubuntu操作系统	用于软件开发和软件部署，每人一台
4	Office、WPS		编写文档

附录3评价标准

评分项目一：实操文档（10分）

表2 实操文档评分细则表

序号	评分项	分值	评分细则
1	实操文档有无	2分	有实操文档得分，无实操文档扣2分。
2	文档任务截图	4分	有操作过程截图得分，无操作过程截图扣4分。

3	文档任务截图标注	4分	有文档任务截图标注说明和画框得分，无标注和画框扣4分

评分项二：依据题的任务，完成任务功能（80分）

表3 项目功能评分细则表

序号	评分项	分值	评分细则
1	功能实现	80分	试题按任务分值评分;根据各任务功能,完成并实现，截图体现关键信息判断，符合要求得该分值，否则不得分。子任务有具体要求,若未完成子任务要求则扣对应子任务3-5分。

评分项三：职业素质（10分）

表4 职业素质评分细则表

序号	评分项	分值	评分细则
1	代码书写格式规范	3分	代码缩进不规范扣1分、方法划分不规范扣1分、语句结构不规范扣1分（如一行编写两个语句）、使用空行不规范扣1分，扣完为止。
2	注释规范	2分	整个项目没有注释扣2分、有注释，但注释不规范扣1分，扣完为止。
3	端口配置	1分	端口号配置不正确，扣1分。
4	部署正确	4分	项目代码未正确配置到指定服务器目录下，扣4分。