# 娄底职业技术学院

Loudi Vocational & Technical College

大数据技术专业
人才培养方案评价资料

# 技能考核题库

娄底职业技术学院
二０二三年八月

# 目录

# 娄底职业技术学院大数据技术专业技能考核题库

本专业技能考核题库包括程序设计、数据库设计、网络爬虫、Hadoop 平台与组件、数据分析，数据可视化 6 个技能考核模块。共 50 题其中程序设计（10 题），数据库设计（5 题），网络爬虫（10 题），Hadoop 平台与组件（20 题），数据可视化（5 题）。用于检测学生的程序设计与开发能力、大数据采集能力、数据存储与处理分析能力、大数据平台部署与运维能力、大数据可视化开发能力以及从事大数据开发工作的程序编写规范、技术文档编写、交流与沟通、法律法规等职业素养。

## 模块 1．专业基本技能模块

### 项目 1 常见算法设计与实现

#### 1.J1-1，《快捷计算系统》关键算法

（1）任务描述

随着网络的不断发展，人们每天接触的新鲜事物都在不断增加，处在这一个信息量大爆炸的时代，时间就尤为重要，为了帮一些人解决时间不充裕的问题，处于创业阶段的某公司准备开发一套“快捷计算”系统，用来解决生活中那些简单、繁琐的数学问题。

任务一：实现平均成绩计算功能的关键算法并绘制流程图（30 分）

已知某个班有 30 个学生，学习 5 门课程，已知所有学生的各科成绩。请编写程序：分别计算每个学生的平均成绩，并输出。

说明：定义一个二维数组 A，用于存放 30 个学生的 5 门成绩。定义一个一维数组 B，用于存放每个学生的 5 门课程的平均成绩。

①使用二重循环，将每个学生的成绩输入到二维数组 A 中。

②使用二重循环，对已经存在于二维数组 A 中的值进行平均分计算，将结

果保存到一维数组 B 中。

③使用循环输出一维数组 B（即平均分）的值。

任务二：实现阶乘计算功能关键算法并绘制流程图（30 分）

输入一个整数 n，计算并输出他的阶乘。

说明：定义一个函数(或方法)，用于求阶乘的值。

在主函数(或主方法)中调用该递归函数(或方法)，求出 5 的阶乘，并输出结果。

任务三：实现前项列和计算功能关键算法并绘制流程图（30 分）

有一分数序列：2/1，3/2，5/3，8/5，13/8，21/13 … 求出这个数列的前 20 项之和。

说明：利用循环计算该数列的和。注意分子分母的变化规律。

a1=2，b1=1，c1=a1/b1；

a2=a1+b1，b2=a1，c2=a2/b2；

a3=a2+b2，b3=a2，c3=a3/b3；

…

s = c1+c2+…+c20；

s 即为分数序列：2/1，3/2，5/3，8/5，13/8，21/13 … 的前 20 项之和。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中，并将所有文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 2.J1-2,《智能统计系统》关键算法

（1）任务描述

生活中在处理各个问题的时候总是会离不开统计，例如统计学生的个数，统计火车买票人数,统计今天是该年的第几天等,所以某团队开发出一套统计系统，用来进行各类统计。

任务一：实现统计今天是该月的有多少天关键算法并绘制流程图（30 分）

从键盘上输入一个年份值和一个月份值,输出该月的天数。(说明:一年有 12 个月，大月的天数是 31，小月的天数是 30。2 月的天数比较特殊，遇到闰年是 29 天，否则为 28 天。

例如，输入 2011、3，则输出 31 天。）

说明：使用分支结构语句实现。

任务二：实现统计纸片对折关键算法并绘制流程图（30 分）

假设一张足够大的纸，纸张的厚度为 0.5 毫米。请问对折多少次以后，可以达到珠穆朗玛峰的高度(最新数据：8844.43 米)。请编写程序输出对折次数。

说明：使用循环结构语句实现，直接输出结果不计分。

任务三：实现统计同构数关键算法并绘制流程图（30 分）

编写程序输出 2~99 之间的同构数。同构数是指这个数为该数平方的尾数，例如 5 的平方为 25，6 的平方为 36，25 的平方为 625，则 5、6、25 都为同构数。

说明：调用带有一个输入参数的函数(或方法)实现，此函数(或方法)用于判断某个整数是否为同构数，输入参数为一个整型参数，返回值为布尔型（是否为同构数)。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中，并将所有文件夹打包压缩，如“娄底职业技术学院

_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 3.J1-3,《儿童网络游戏游戏》关键算法

（1）任务描述

A 公司是专门的儿童网络游戏公司，现在公司正在开发几款智力游戏，其中需要设计几个算法模型。

任务一：实现堆积木游戏功能关键算法并绘制流程图（30 分）

堆积木是小孩子最爱玩的游戏,但是因为小孩子的好奇心(比如误食积木等)导致家长们越不愿意让孩子去玩积木，为了解决这个问题 TX 公司开发了一套 VR 积木游戏，你要做的是将用户堆好的积木在屏幕中显示出来。

*******

*****

***

*

注意：使用循环结构语句实现。

任务二：实现抓娃娃游戏功能关键算法并绘制流程图（30 分）

请你在娃娃机里放十个娃娃，每个娃娃对应一个数字，该数字表示娃娃的大小。要求通过计算能输出最大的娃娃对应的数字，你可以这样做：

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。

注意：使用数组、循环结构语句实现。

任务三：实现算数游戏功能关键算法并绘制流程图（30 分）

游戏主要是这样的，计算正整数 n 每个数位上的数之积，例如 24，它的每个数位上的数字之积为 2 * 4 = 8，现在要求你为 A 公司编写一个计算函数(或方法)fun，将结果放到 c 中，并显示输出。作为参考答案。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中，并将所有文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 4.J1-4,《英语辅导系统》关键算法

（1）任务描述

随着国际化的到来英语在人们生活中就凸显得比较重要了,特别是学习编程语言的人们，所以 B 公司决定开发一套英语辅助学习系统，通过完成趣味试题，采用游戏通关的方式，帮助有需要的人更好的学习英语。

任务一：实现趣味英语试题 1 关键算法并绘制流程图（30 分）

已知字符串数组 A，包含初始数据：a1,a2,a3,a4,a5；字符串数组 B，包含初始数据：b1,b2,b3,b4,b5。编写程序将数组 A、B 的每一对应数据项相连接，然后存入字符串数组 C，并输出数组 C。输出结果为：a1b1,a2b2,a3b3,a4b4,a5b5。

例如：数组 A 的值为{“Hello ”，“Hello ”，“Hello ”，“Hello ”，

“Hello ” }，数组 B 的值为{“Jack” ，“Tom” ，“Lee” ，“John” ，“Alisa” }，则输出结果为{“Hello Jack” ，“Hello Tom” ，“Hello Lee”，“Hello John” ，“Hello Alisa” }。

注意：定义 2 个字符串数组 A、B，用于存储读取数据。定义数组 C，用于输出结果。

①使用循环将数组 A、B 的对应项相连接，结果存入数组 C。

②使用循环将数组 C 中的值输出。

任务二：实现趣味英语试题 2 关键算法并绘制流程图（30 分）。

判断一个字符串是否是对称字符串，例如:“abc”不是对称字符串,“aba”、“abba”、“aaa”、“mnanm”是对称字符串。是的话输出“Yes”，否则输出“No”。

注意：使用循环和判断语句实现。

任务三：实现趣味英语试题 3 关键算法并绘制流程图（30 分）

编写一个程序实现统计一串字符串中的英文小写字母个数！例如：输入“axZvnNgOuyi”，得到的值应该是 8;

注意：使用分支语句实现，且有输入输出，直接输出不计分。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中，并将所有文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 5.J1-5,《手机号码查询系统》关键算法

（1）任务描述

现在手机使用非常普及，为方便人们查询手机号码的归属地信息，A 公司决定开发一个手机号码查询系统，需要完成以下任务。

任务一：实现手机号计数功能关键算法并绘制流程图（30 分）

从键盘接收一行字符串，字符串中只包含数字和空格，统计其中所有的手机号码数量。比如输入：18711389426 18711389427 输出的结果为：2。

注意：使用分支及循环结构完成。

任务二：实现连号判断功能关键算法并绘制流程图（30 分）。

从键盘接收一个十一位的数字，判断其是否为尾号 5 连的手机号。规则：第 1 位是 1,第二位可以是数字 358 其中之一,后面 4 位任意数字,最后 5 位为任意相同的数字。例如：18601088888、13912366666 则满足。

注意：不满足的输出“false”，满足要求的输出“true”。

任务三：实现统计非数字功能关键算法并绘制流程图（30 分）

对于给定的一个字符串，统计其中非数字字符出现的次数。

例如：输入：Ab(&%123) 输出：6

注意：使用循环和判断语句实现。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中,并将所有文件夹打包压缩,如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 6.J1-6,《趣味数学学习系统》关键算法

（1）任务描述

由于中学数学是培养数学思维的基础阶段,为了让学生打造一个坚实的数学基础，A 学校决定开发一个趣味数学学习系统，通过完成趣味试题，采用游戏通关的方式，帮助中学生初步掌握二元一次方程解简单应用题的方法和步骤，并会列出二元一次方程解简单的应用题。

任务一：实现汽车与摩托问题的关键算法并绘制流程图（30 分）

在一个停车场内，汽车、摩托车共停了 48 辆，其中每辆汽车有 4 个轮子，每辆摩托车有 3 个轮子，这些车共有 172 个轮子，编程输出停车场内有汽 车和摩托车的数量。

注意：用循环语句实现。

任务二：实现鸡兔同笼问题的关键算法并绘制流程图（30 分）。

已知鸡和兔的总数量为 n,总腿数为 m。输入 n 和 m,依次输出鸡和兔的数目，如果无解，则输出“No answer”(不要引号)。

注意：用循环语句实现。

任务三：实现合格电视机问题的关键算法并绘制流程图（30 分）

某电视机厂每天生产电视 500 台，在质量评比中，每生产一台合格电视机记 5 分，每生产一台不合格电视机扣 18 分。如果四天得了 9931 分，编程计算这四天生产的合格电视机的台数，并输出。

注意：用循环语句实现。。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”

命名的总文件夹中,并将所有文件夹打包压缩,如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 7.J1-7,《在线评判系统》题库关键算法

（1）任务描述

在线评判系统(简称 OJ,Online Judge)指在线用来评判程序的正确性、时间与效率空间效率的评判系统。现需要为特定题目设计正确算法以便扩充题库，请完成以下任务。

任务一：实现问题一关键算法并绘制流程图（30 分）

编写一个程序，该程序读取一个字符串，然后输出读取的空格数目。

注意：输入字符串的长度不超过 30 个字符（含空格）。

任务二：实现问题二关键算法并绘制流程图（30 分）。

中国古代的《算经》记载了这样一个问题：公鸡 5 文钱 1 只，母鸡 3 文钱 1 只，小鸡 1 文钱 3 只，如果用 100 文钱买 100 只鸡，那么公鸡、母鸡和小鸡各应该买多少只呢？现请你编程求出所有的解，每个解输出 3 个整数，打印在一行，用空格隔开，分别代表买的公鸡、母鸡、小鸡的数量。

注意：100 文钱要正好用完。请输出所有的解，每个解占一行。

任务三：实现问题三关键算法并绘制流程图（30 分）

有一天爱因斯坦给他的朋友出了一个题目，有一个楼，其两层之间有一个很长的阶梯。如果一个人每步上 2 阶，最后剩 1 阶；如果一个人每步上 3 阶，

最后剩 2 阶；如果一个人每步上 5 阶，最后剩 4 阶；如果一个人每步上 6 阶，最后剩 5 阶；如果一个人每步上 7 阶，后刚好一阶也不剩。问这个阶梯至少有多少阶呢？

注意：请编程求出最小的一个答案并输出。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中,并将所有文件夹打包压缩,如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 8.J1-8,《生活繁琐计算系统 》关键算法

（1）任务描述

随着我国经济的发展,社会的进步,交易额每天都在不断上升,所以在人们生活中的各种计算问题不断显现出来，例如税收、比赛评分等问题的计算，当数据多了难免会出问题，所以开发出一套这种系统存在着一定的意义。

任务一：实现评分计算功能关键算法并绘制流程图（30 分）

编写一个应用程序，计算并输出一维数组（9.8，12，45，67，23，1.98，2.55，45）中的最大值、最小值和平均值。

任务二：实现规律数字计算关键算法并绘制流程图（30 分）。

计算算式 1+21+22+23+…+2n 的值。

注意：n 由键盘输入，且 2 ≤ n ≤10。

任务三：实现个人交税计算功能关键算法并绘制流程图（30 分）

某国的个人所得税草案规定，个税的起征点为 3000 元，分成 7 级，税率情况见下表，从键盘上输入月工资，计算应交纳的个人所得税。

表 1.6.1 税率情况表

级数　全月应纳税所得额　税率 (%)

1　不超过 1500 元的（即 3000-4500 之间）　5

2　超过 1500 元至 4500 元的部分　10

3　超过 4500 元至 9000 元的部分　20

4　超过 9000 元至 35000 元的部分　25

5　超过 35000 元至 55000 元的部分　30

6　超过 55000 元至 80000 元的部分　35

7　超过 80000 元的部分　45

注意：超出部分按所在税的级数计算，如：一个人的月收入为 6000，应交个人所得税为：1500*0.05 +((6000-3000）-1500)*0.1=225

请在键盘上输入一个人的月收入，编程实现计算该公民所要交的税。

例如：输入“6000”，则输出“你要交的税为：225”。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中,并将所有文件夹打包压缩,如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 9.J1-9,《成绩分析系统》关键算法

（1）任务描述

对学生成绩进行统计和数据分析可以发现学生对知识的掌握情况,以便教师根据分析的结果调整教学内容和重难点,现在需要完成以下任务来实现成绩分析系统。

任务一：实现成绩等级划分功能关键算法并绘制流程图（30 分）

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 A

80~89 为 B

70~79 为 C

60~69 为 D

0~59 为 E

要求：如果输入数据不在 0~100 范围内，请输出一行:“Score is error!”。

任务二：实现数列求和功能关键算法并绘制流程图（30 分）。

数列的定义如下：

数列的第一项为 n，以后各项为前一项的平方根，输出数列的前 m 项的和。

要求：数列的各项均为正数。

任务三：求前 n 项之和功能关键算法并绘制流程图（30 分）

多项式的描述如下：1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + ...，现在要求出该多项式

的前 n 项的和。

要求：结果保留两位小数。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中,并将所有文件夹打包压缩,如“娄底职业技术学院_43******_

张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 1：算法设计与实现评分标准

（4）实施条件

见附录 4：程序设计模块实施条件

## 10. J1-10,《 市场分析系统 》 关键算法

（1）任务描述

在一个新的产品要上市的之前，需要做大量的市场调查，以确保产品能获得理想的收益。现在 A 公司要设计一款市场分析系统，需完成以下功能模块。

任务一：实现销售分析功能关键算法并绘制流程图（30 分）

A 商店准备在今年夏天开始出售西瓜，西瓜的售价如下，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。现在为了知道商店是否会盈利要求 A 公司帮忙设计一个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的程序。

注意：使用分支结构语句实现，结果保留两位小数。

任务二：实现销售量分析功能关键算法并绘制流程图并绘制流程图（30 分）

KJ 学院为全校同学设计一套校服，A 公司有意招标为 A 学校设计服装，职员小 C 在 A 校排队时偷偷的看了一眼发现 A 学校学生，5 人一行余 2 人，7 人一行余 3 人，3 人一行余 1 人，编写一个程序求该校的学生人数。

注意：使用分支、循环结构语句实现，直接输出结果不计分。

任务三：实现市场调查数据的恢复功能关键算法并绘制流程图（30 分）

职员小 A 今天犯了一个致命的错误,他一不小心丢失了 X 项目的市场调查结果只记得一个公式 xyz+yzz=532，其中 x、y、z 均为一位数，现在请你帮忙

编写一个程序求出 x、y、z 分别代表什么数。

注意：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。

作品提交：创建”源代码”文件夹将三个任务的源代码保存到其中，并将程序运行结果截图保存至”结果.doc”文档中,创建“所属学校_身份证_姓名_题号”命名的总文件夹中,并将所有文件夹打包压缩,如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。(2) 考核时量

考核时间为 2 个小时。

(3) 评分标准

见附录 2：数据库设计评分标准

(4) 实施条件

见附录 4：数据库设计模块实施条件

# 项目 2 MySql 数据库操作与查询

## 1.J2-1，人力资源管理-人员管理数据库设计 1

(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 2.1.1 所示，物理数据模型如图 2.1.2 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

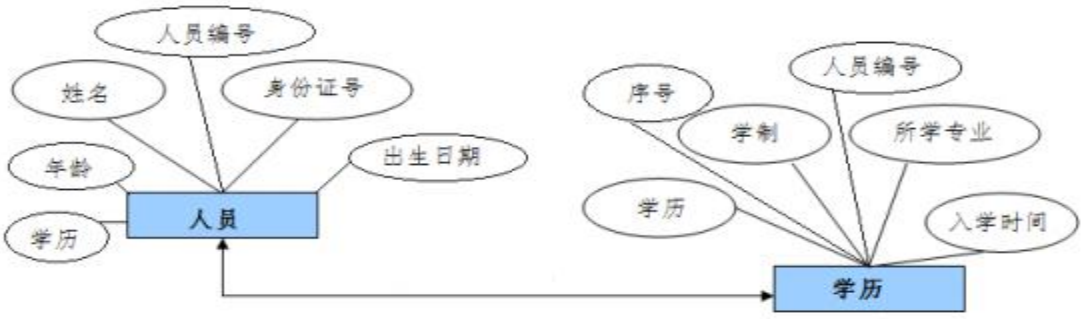


图 2.1.1 E-R 图


图 2.1.2 物理数据原型

表 2.1.1 字段名定义

| 表 t_staff | | 表 t_educational | |
|---|---|---|---|
| 字段名 | 字段说明 | 字段名 | 字段说明 |
| staff_no | 编号 | id | 编号（主键自增） |
| name | 姓名 | degree | 学历 |
| ic_card | 身份证号 | major | 专业 |
| age | 年龄 | reg_time | 入学时间 |
| bithday | 生日 | length_of_schooling | 学制 |
| | | staff_no | 人员编号 |

操作步骤：

① 创建数据库 resourcesDB (5 分)。

② 根据 ER 图，物理数据原型图，字段表完成表 t_staff,t_educational 的创建(20 分)。

③ 根据物理数据原型设置数据关系(10 分)。

④ 使用 SQL 完成如下操作(55 分)。

1) 向每个表中插入 5 条测试数据(10 分)。

2) 更新 t_staff 表，将所有“李”姓变成“王”姓(10 分)。

3) 查询所有年龄大于 16 的人员信息(10 分)。

4) 查询出拥有“大学本科”学历的所有人员姓名(10 分)。

5) 查询出学历不是“大学本科”的所有人员姓名，年龄，按照年龄降序排序(15 分)。

作品提交：提交数据库脚本文件 resourcesDB.sql，以及数据库操作 SQL 文件 操作.sql;再将 2 个文件保存到“所属学校_身份证_姓名_题号”命名的总文

件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”，将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 2：数据库设计评分标准

（4）实施条件

见附录 4：数据库设计模块实施条件

## 2.J2-2，人力资源管理-人员管理数据库设计 2

（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 2.2.1 所示，物理数据模型如图 2.2.2 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 2.2.1 E-R 图

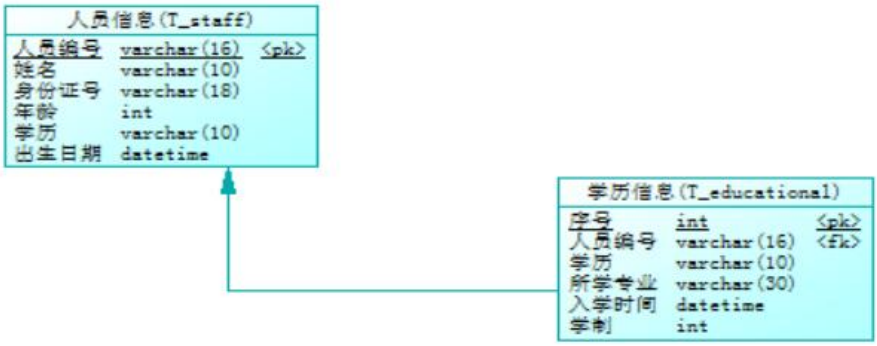


图 2.2.2 物理数据原型

表 2.2.1 字段名定义

| 表 t_staff | | 表 t_educational | |
|---|---|---|---|
| 字段名 | 字段说明 | 字段名 | 字段说明 |
| staff_no | 编号 | id | 编号（主键自增） |
| name | 姓名 | degree | 学历 |
| ic_card | 身份证号 | major | 专业 |
| age | 年龄 | reg_time | 入学时间 |
| bithday | 生日 | length_of_schooling | 学制 |
| | | staff_no | 人员编号 |

操作步骤：

① 创建数据库 resourcesDB(5 分)。

② 根据 ER 图，物理数据原型图，字段表完成表 t_staff,t_educational 的创建(20 分)。

③ 根据物理数据原型设置数据关系(10 分)。

④ 使用 SQL 完成如下操作(55 分)。

1) 向每个表中插入 5 条测试数据(10 分)。

2) 将所有人员年龄都增加 1 岁(10 分)。

3) 查询出 t_staff 表中大于平均年龄的人员名单(10 分)。

4) 查询出学习“软件专业”所有人员姓名，年龄(10 分)。

5) 查询出还未毕业的所有人员姓名，年龄，按照年龄降序排序(15 分)。

作品提交：提交数据库脚本文件 resourcesDB.sql，以及数据库操作 SQL 文件 操作.sql;再将 2 个文件保存到“所属学校_身份证_姓名_题号”命名的总文

件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”，将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 2：数据库设计评分标准

（4）实施条件

见附录 4：数据库设计模块实施条件

## 3. J2-3，人力资源管理-员工工资管理数据库设计

（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 2.3.1 所示，物理数据模型如图 2.3.2 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

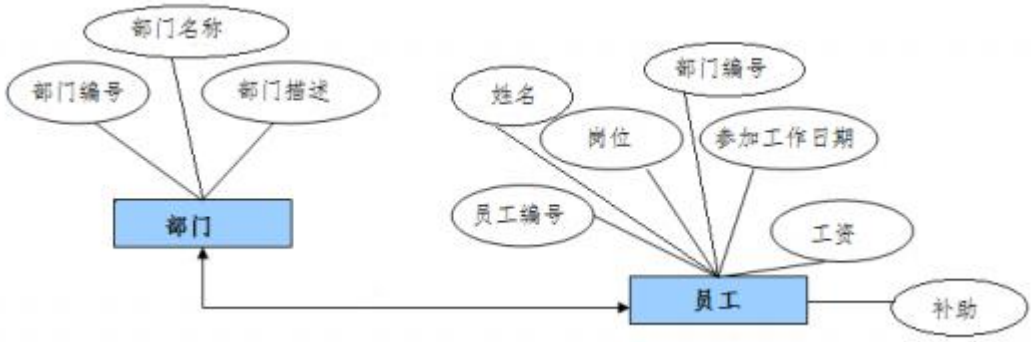


图 2.3.1 E-R 图

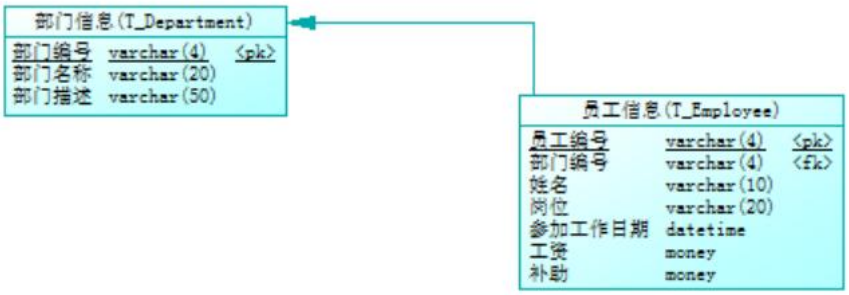


图 2.3.2 物理数据原型

表 2.3.1 字段名定义

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

操作步骤：

① 创建数据库 salaryDB(5 分)。

② 根据 ER 图，物理数据原型图，字段表完成表 t_department,t_employee 的创建(20 分)。

③ 根据物理数据原型设置数据关系(10 分)。

④ 使用 SQL 完成如下操作(55 分)：

1) 向每个表中插入 5 条测试数据(10 分)。

2) 将所有人工资上浮 10%(10 分)。

3) 查询所有部门编号为“d001”的员工姓名，岗位，参加工作时间及工资(10 分)。

4) 查询出岗位是开发的平均工资(10 分)。

5) 查询每个员工的年薪，员工姓名，部门名称，并按照年薪降序排列(15 分)。

作品提交:提交数据库脚本文件 salaryDB.sql,以及数据库操作 SQL 文件 操作.sql;再将 2 个文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 2：数据库设计评分标准

（4）实施条件

见附录 4：数据库设计模块实施条件

## 4.J2-4，建设用地信息系统-基础数据设置系统数据库设计

（1）任务描述

《建设用地信息系统》基础数据设置子模块的 E-R 图如图 2.4.1 所示，物理数据模型如图 2.4.2 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

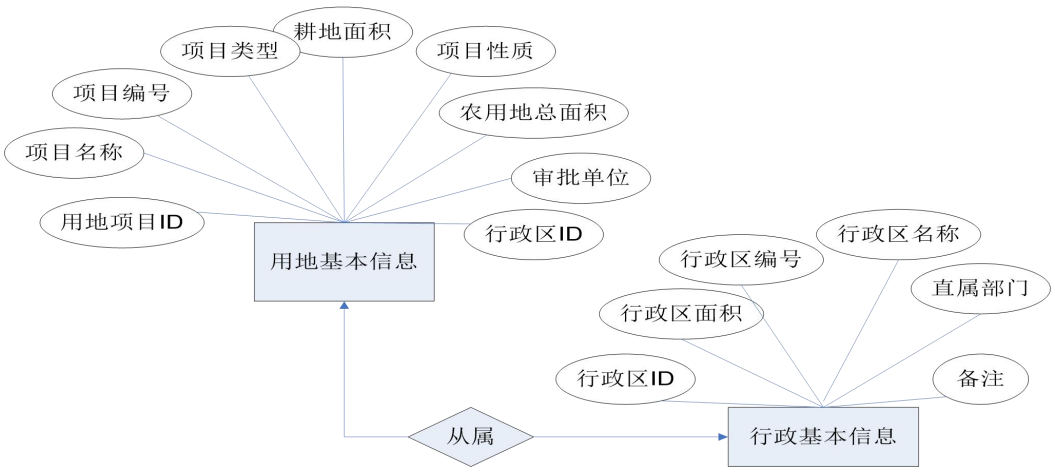


图 2.4.1 E-R 图

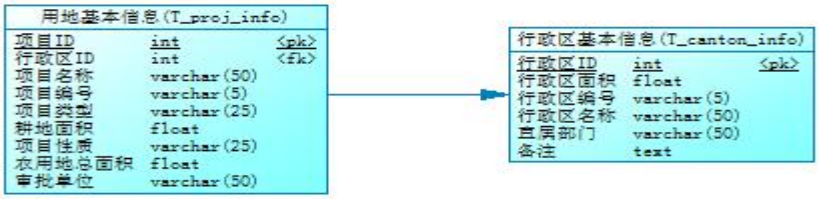


图 2.4.2 物理数据原型

表 2.4.1 字段名定义

| 表 t_canton_info | | 表 t_proj_info | |
|---|---|---|---|
| 字段名 | 字段说明 | 字段名 | 字段说明 |
| canton_id(标识列) | 行政区 ID | proj_id(标识列) | 用地项目 ID |
| canton_no | 行政区编号 | proj_no | 项目编号 |
| canton_name | 行政区名称 | proj_name | 项目名称 |
| canton_tot | 行政区面积 | proj_type | 项目类型 |
| branch | 直属部门 | proj_kind | 项目性质 |
| remark | 备注 | farm_tot | 农用地总面积 |
| | | tilth_state | 耕地面积 |
| | | approve_unit | 审批单位 |
| | | canton_id | 行政区编号 |

操作步骤：

① 创建数据库 areaprjectDB(5 分)。

② 根据 ER 图，物理数据原型图，字段表完成表 t_proj_info，t_canton_info 的创建(20 分)。

③ 根据物理数据原型设置数据关系(10 分)。

④ 使用 SQL 完成如下操作(55 分)：

1) 向每个表中插入 5 条测试数据(10 分)。

2) 将直属部门由“长沙市国土资源局”修改为“株洲市国土资源局”(10 分)。

3) 查询出项目编号为 C0001 的建设用地基本信息(10 分)。

4) 查询出所有的建设土地基本信息并按农用地总面积升序排序(10 分)。

5) 查询出行政直属部门为“长沙市国土资源局”的建设用地基本信息(15分)。

作品提交：提交数据库脚本文件 areaprjectDB.sql，以及数据库操作 SQL 文件 操作.sql;再将 2 个文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 2：数据库设计评分标准

（4）实施条件

见附录 4：数据库设计模块实施条件

## 5.J2-5，建设用地信息系统-报批管理系统数据库设计

（1）任务描述

《建设用地信息系统》报批管理子模块的 E-R 图如图 2.5.1 所示，物理数据模型如图 2.5.2 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

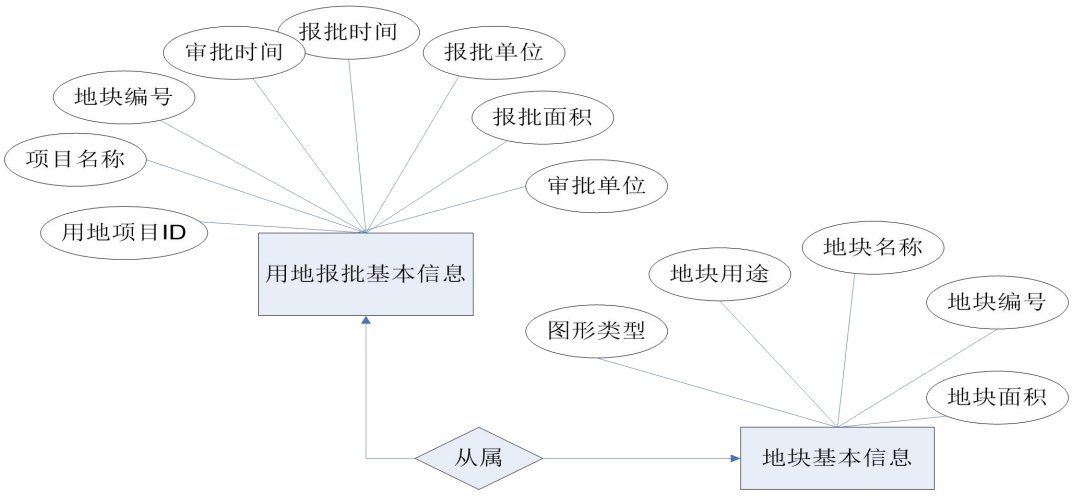

图 2.5.1 E-R 图

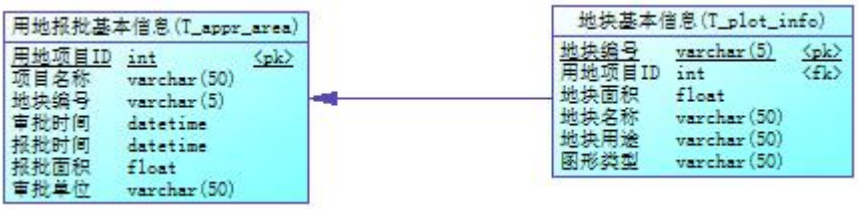


图 2.5.2 物理数据原型

表 2.5.1 字段名定义

| 表 t_appr_area | | 表 t_plot_info | |
|---|---|---|---|
| 字段名 | 字段说明 | 字段名 | 字段说明 |
| proj_id | 用地项目 ID | plot_id | 地块编号 |
| proj_name | 项目名称 | plot_name | 地块名称 |
| plot_id | 地块编号 | total_area | 地块面积 |
| appr_area | 报批面积 | purpose | 地块用途 |
| appr_date | 报批时间 | shape_type | 图形类型 |
| appro_unit | 审批单位 | proj_id | 用地项目 ID |
| appro_date | 审批时间 | | |

操作步骤：

① 创建数据库 newdatasetDB(5 分)。

② 根据 ER 图，物理数据原型图，字段表完成表 t_appr_area,t_plot_info 的创建(20 分)。

③ 根据物理数据原型设置数据关系(10 分)。

④ 使用 SQL 完成如下操作(55 分)：

1) 向每个表中插入 5 条测试数据(10 分)。

2) 将地块名称为“长沙市天心花苑”修改为“株洲市天心花苑”(10 分)。

3) 查询地块编号为“10001”的建设用地的报批基本信息(10 分)。

4) 查询项目“保利一期”项目中所有地块信息(10 分)。

5) 查询所有的地块基本信息，包括所属项目名称，并按地块面积升序排序

(15 分)。

作品提交：提交数据库脚本文件 newdatasetDB.sql，以及数据库操作 SQL 文件 操作.sql;再将 2 个文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 2 个小时。

（3）评分标准

见附录 2：数据库设计评分标准

（4）实施条件

见附录 4：数据库设计模块实施条件

# 模块 2. 岗位核心技能模块

## 项目 3　网络爬虫与分析

### 1.H1-1，期货网站信息爬取

（1）任务描述

基于 Python 爬虫代码从指定网站抓取所有期货商品的报价，爬取内容包括：商品名称，规格，最新报价。 网站：

http://price.mofcom.gov.cn/pricequotation/morepricequotation.shtml?flag=qh&prod_type=lntx


图 3.1.1 网站例图

实施步骤：

① 分析该网页代码，获取对应 header 与请求 url。

② 正确导入 urllib 等库。

③ 通过网页分析获取头部信息以及网页结构和网络请求。

④ 定义 getgoods()函数，使用 requests 或其他方法获取期货第一页信息。

⑤ 通过解析定义方法 getprice()函数，使用 requests 或其他方法获取每个期货最近期货收盘价。

⑥ 使用 pymysql 或其他组件保存商品信息到表 tbgoods 中。

| 字段名 | 说明 |
| --- | --- |

| | |
|---|---|
| goodsId | 商品 id |
| goodsName | 商品名称 |
| goodsType | 规格信息 |

⑦ 使用 pymysql 或其他组件保存价格到到表 tbprices 中。

| 字段名 | 说明 |
|---|---|
| goodsId | 商品 id |
| times | 交易时间 |
| price | 报价 |

作品提交：将项目文件以及数据库文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”，将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 5 分 | 能够正确创建 python 爬虫项目 |
| | 导入库 | 5 分 | 正确导入 requests，xpath 相关库 |
| | 获取期货信息 | 25 分 | 1、正确定义函数 getgoods 5 分<br>2、请求 url 设置正确 5 分<br>3、正确使用 requests 等完成接口请求得到信息 10 分<br>4、通过处理得到对应期货信息列表 5 分 |
| | 获取期货价格 | 40 分 | 1、正确定义函数 getprice 5 分<br>2、请求 url 设置正确 5 分<br>3、正确使用 requests 等完成接口请求 |

| | | | |
|---|---|---|---|
| | | | 得到价格 10 分<br>4、通过处理得到对应价格列表 5 分 |
| | 保存数据 | 15 分 | 1. 正确设置数据库结构 10 分<br>2. 正确存储数据到数据库,表 1 数据正确:10 分,表 2 数据正确:10 分 |
| 素养要求 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 2. H1-2，天气数据信息爬取

（1）任务描述

基于 Python 爬虫代码从指定天气数据查询网站抓取算法数据，爬取内容包括：略，天气页面上有提示[所有有红色菱形的城市都需要获取]。

网站：http://www.envicloud.cn/dataMap?title=3

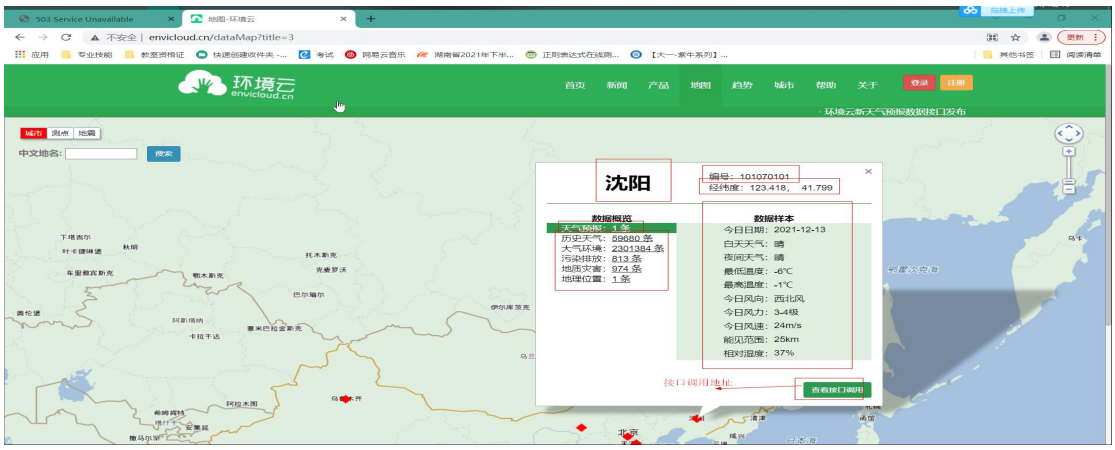


图 3.2.1 网站例图

实施步骤：

① 正确创建爬虫程序（5 分）。

② 正确导入 urllib 等库(5 分)。

③ 通过分析网络请求得到红点请求 url，详情请求 url。

④ 定义 getdots()函数，使用 requests 得到所有红点信息，数据存储到集合中(25 分)。

⑤ 定义 getdetail()函数，遍历所有红点获取详细信息（地点，经纬度，天气预报，历史天气，大气环境，污染排放，地质灾害，地理位置），并获取该点天气预报详情(40 分)。

⑥ 对详细信息逐条存储到 D://天气.csv 文件中，期中天气预报详情作为一个字符串保存(15 分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 5 分 | 能够正确创建 python 爬虫项目 |
| | 导入库 | 5 分 | 正确导入 urllib 相关库 |
| | 红点数据获取 | 25 分 | 5、正确定义函数 getdots 5 分<br>6、红点请求 url 设置正确 5 分<br>7、正确使用 requests 等完成接口请求得到红点信息 10 分<br>8、通过处理保存红点数据到集合中 5 分 |
| | 详情信息获取 | 40 分 | 1、正确定义函数 getdetail 5 分<br>2、详情请求 url 设置正确 5 分<br>3、正确遍历所有红点信息获取详情信息：地点，经纬度，天气预报，历史天气，大气环境，污染排放，地质灾害， |

| | | | |
|---|---|---|---|
| | | | 地理位置　每点 2 分 总分：16 分<br>4、获取该点天气预报详情 14 分 |
| | 保存数据 | 15 分 | 能够把地点，经纬度，天气预报，历史天气，大气环境，污染排放，地质灾害，地理位置 ,天气预报详情保存对应 csv 文件中。15 分 |
| 素养要求 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 3.H1-3，2022 畅销书籍爬取

（1）任务描述

基于 Python 爬虫代码从指定读书网站抓取畅销榜数据前 10 页，爬取内容包括：书籍图片，书籍名称，书籍折扣，书籍价格。

网址：

http://bang.dangdang.com/books/bestsellers/01.00.00.00.00.00-24hours-0-0-1-1


图 3.3.1 网站例图

实施步骤：

① 正确创建爬虫程序(5 分)。

② 正确导入 urllib 等库(5 分)。

③ 通过分析得到正确头部信息，以及网页构造(20 分)。

④ 定义 gethtml()函数，通过 urllib 库中对应方法获取该网页信息(20 分)。

⑤ 创建解析函数 parsehtml(html)用于处理解析，使用 Beautiful Soup 获取到书籍图片，书籍名称，书籍折扣，书籍作者，书籍价格(35 分)。

⑥ 创建 nextpage()函数爬取下一页信息直到爬取前 10 页结束停止爬取(10 分)。

⑦ 把获取到的书籍名称，书籍折扣，书籍作者，书籍价格保存到 D://书籍列表.csv 文件中，书籍图片保存到 D://bookimgs 文件夹中(20 分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_

张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 5 分 | 能够正确创建 python 爬虫项目 |
| | 导入库 | 5 分 | 正确导入 urllib 相关库 |
| | 网页爬取 | 20 分 | 1. 正确创建 gethtml 函数 5 分<br>2. 正确创建 requests 对象 5 分<br>3. 正确设置 header 属性 5 分<br>4. 正确获取对应 html 数据并返回 5 分 |
| | 数据分析 | 35 分 | 1、正确定义函数 parsehtml(html) 5 分<br>2、正确使用 Beutiful Soup 转换 html 为对象 5 分<br>3、使用 Bs4 对应方法获取书籍图片路径，书籍名称，书籍折扣，书籍作者，书籍价格 每项 5 分 总分： 25 分 |
| | 数据保存 | 30 分 | 1、下截图片保存到 D://bookimgs 文件夹下 15 分<br>2、把书籍名称，书籍折扣，书籍作者，书籍价格保存到对应 csv 文件中 15 分 |
| 素养要求 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 4.H1-4，畅销书籍评论爬取

（1）任务描述

基于 Python 爬虫代码从畅销榜第一页所有图书评论,内容包括：评论作者，评论内容，评分，评论标题，评论日期。

网站

http://bang.dangdang.com/books/bestsellers/01.00.00.00.00.00-24hours-0-0-1-1


图 3.4.1 网站例图

实施步骤：

① 正确创建爬虫程序(5 分)。

② 正确导入 urllib 等库(5 分)。

③ 通过分析得到正确头部信息，以及网页构造(20 分)。

④ 定义 gethtml()函数，通过 urllib 库中对应方法获取该网页信息(35 分)。

⑤ 创建解析函数 parsehtml(html)用于处理解析，使用 Beautiful Soup 或者 xpath 获取到评论作者，评论内容，评分，评论标题，评论日期(15 分)。

⑥ 创建 nextnotice()函数爬取下一个图书评论信息直到爬取完第一页所有书籍评论停止爬取(15 分)。

⑦ 将获取到的评论作者，评论内容，评分，评论标题，评论日期保存到 D://书籍评论.csv 中(15 分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 5 分 | 能够正确创建 python 爬虫项目 |
| | 导入库 | 5 分 | 正确导入 urllib 相关库 |
| | 网页爬取 | 20 分 | 1. 正确创建 gethtml 函数 5 分<br>2. 正确创建 requests 对象 5 分<br>3. 正确设置 header 属性 5 分<br>4. 正确获取对应 html 数据并返回 5 分 |
| | 数据分析 | 35 分 | 1、正确定义函数 parsehtml(html) 5 分<br>2、正确完成 Beutiful Soup 或 xpath 的初始化 5 分<br>3、使用 Bs4 对应方法获取评论作者，评论内容，评分，评论标题，评论日期 每项 5 分 总分： 25 分 |
| | 下一页功能 | 15 分 | 1、正确定义函数 next() 5 分<br>2、完成下一页逻辑 10 分 |
| | 保存 | 15 分 | 1、将获取到的评论作者，评论内容，评分，评论标题，评论日期保存到 D://书籍评论.csv 中 15 分 |
| 素养要求 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 5.H1-5，网易新闻信息爬取

（1）任务描述

基于 Python 爬虫代码指定页面爬取新闻图片，新闻标题，新闻时间，新闻关键词，跟帖数。

网站：https://news.163.com/domestic/


图 3.5.1 网站例图

实施步骤：

① 正确创建爬虫程序(5 分)。

② 正确导入 urllib 等库(5 分)。

③ 通过分析得到正确头部信息，以及网页构造(20 分)。

④ 定义 gethtml()函数，通过 urllib 库中对应方法获取该网页信息(35 分)。

⑤ 创建解析函数 parsehtml(html)用于处理解析，使用 Beautiful Soup 获取到新闻图片地址，新闻标题，新闻时间，新闻关键词，跟帖数(35 分)。

⑥ 把获取到的新闻标题，新闻时间，新闻关键词，跟帖数保存到D://新闻列表.csv文件中，新闻图片地址保存到D://news 文件夹中(30分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为3个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 5分 | 能够正确创建python爬虫项目 |
| | 导入库 | 5分 | 正确导入urllib相关库 |
| | 网页爬取 | 20分 | 1. 正确创建gethtml函数 5分<br>2. 正确创建requests对象 5分<br>3. 正确设置header属性 5分<br>4. 正确获取对应html数据并返回5分 |
| | 数据分析 | 35分 | 1、正确定义函数parsehtml(html) 5分<br>2、正确使用Beutiful Soup 转换html为对象 5分<br>3、使用Bs4对应方法获取新闻图片地址，新闻标题，新闻时间，新闻关键词，跟帖数 每项5分 总分： 25分 |
| | 数据保存 | 30分 | 1、下截图片保存到D://news文件夹下15分<br>2、把新闻标题,新闻时间,新闻关键词,跟帖数保存到对应D://新闻.csv文件中15分 |
| 素养要求 | 代码书写规范 | 3分 | 代码缩进不规范扣1分、方法定义不规范扣1分、语句结构不规范扣1分 |
| | 注释规范 | 2分 | 无注释扣2分，注释不规范扣1分 |

| | | | |
|---|---|---|---|
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 6. H1-6，招聘网站信息爬取

（1）任务描述

基于 Python 爬虫代码从指定人才网数据查询网站抓取算法数据，爬取内容包括：岗位信息，地区，工作，经验，学历，福利。

网站：http://www.pjob.net/china.htm


图 3.6.1 网站例图

实施步骤：

① 正确创建爬虫程序(5 分)。

② 正确导入 urllib 等库(5 分)。

③ 通过分析得到正确头部信息，以及网页构造。

④ 定义 gethtml()函数，通过 urllib 库中对应方法获取该网页信息(20 分)。

⑤ 创建解析函数 parsehtml(html)用于处理解析，使用 Beautiful Soup 获取到岗位名称，地区，工作，经验，学历，福利(40 分)。

⑥ 把获取到的岗位名称，地区，工作，经验，学历，福利（福利信息每项用#分隔）保存到D://新闻列表.csv文件中(25分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为3个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 创建项目 | 5分 | 能够正确创建python爬虫项目 |
| | 导入库 | 5分 | 正确导入urllib相关库 |
| | 网页爬取 | 20分 | 1. 正确创建gethtml函数 5分<br>2. 正确创建requests对象 5分<br>3. 正确设置header属性 5分<br>4. 正确获取对应html数据并返回5分 |
| | 数据分析 | 40分 | 1、正确定义函数parsehtml(html) 5分<br>2、正确使用Beutiful Soup 转换html为对象 5分<br>3、使用Bs4对应方法获取岗位信息，地区，工作，经验，学历，福利 每项5分 总分： 30分 |
| | 数据保存 | 25分 | 1、格式化福利信息 5分<br>2、把岗位信息，地区，工作，经验，学历，福利保存到对应D://新闻.csv文件中 20分 |
| 素养要求 | 代码书写规范 | 3分 | 代码缩进不规范扣1分、方法定义不规范扣1分、语句结构不规范扣1分 |
| | 注释规范 | 2分 | 无注释扣2分，注释不规范扣1分 |

| | | | |
|---|---|---|---|
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 7. H1-7，中国福布斯排行榜爬取

（1）任务描述

基于 Python 爬虫代码从福布斯排行榜爬取内容包括：排名，名字，财富，来源，年龄，城市。

网站：https://www.maigoo.com/news/572609.html


| 排名 | 姓名 | 财富（亿元） | 财富来源 | 年龄 | 居住城市 |
|---|---|---|---|---|---|
| 11 | 丁磊 | 1781.6 | 网易 | 49 | 杭州 |
| 12 | 张勇家族 | 1748.2 | 海底捞 | / | 新加坡 |
| 13 | 庞康 | 1661.5 | 海天味业 | 64 | 佛山 |
| 14 | 秦英林家族 | 1474.6 | 牧原股份 | 55 | 南阳 |
| 15 | 左晖 | 1374.5 | 贝壳 | 49 | 北京 |
| 16 | 王兴 | 1367.9 | 美团点评 | 41 | 北京 |
| 17 | 刘强东 | 1354.5 | 京东 | 46 | 北京 |
| 17 | 雷军 | 1354.5 | 小米集团 | 50 | 北京 |
| 19 | 曾毓群 | 1341.2 | 宁德时代 | 51 | 宁德 |
| 20 | 李西廷 | 1301.1 | 迈瑞医疗 | 69 | 深圳 |
| 20 | 张志东 | 1301.1 | 腾讯 | 48 | 深圳 |

图 3.7.1 网站例图

实施步骤：

① 正确创建 scrapy 爬虫程序(10 分)。

② 正确配置 settings 文件（头部，缓存，超时等）(10 分)。

③ 正确定义 items 文件(5 分)。

④ 完成数据提取 spider 使用 xpath 或者 BS4 等完成：排名，名字，财富，来源，年龄，城市 这几项数据的获取(50 分)。

⑤ 在 pipeline 中对数据进行存储，存储到 D://福布斯.csv 文件中(20 分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 10 分 | 能够正确创建 scrapy 爬虫项目 |
| | Settings 配置 | 10 分 | 1. 正确配置头部信息 5 分<br>2. 正确配置超时等基本信息 5 分 |
| | 定义 items | 5 分 | 1.根据查找数据定义对应 items 5 分 |
| | 数据提取 | 50 分 | 1. 正确创建 spider 5 分<br>2. 初始化 xpath 或 bs4 5 分<br>3. 完成排名，名字，财富，来源，年龄，城市这几项数据的获取 30 分<br>4. 正确转换数据为 items 10 分 |
| | 保存数据 | 20 分 | 1.在 pipeline 中对数据排名，名字，财富，来源，年龄，城市进行存储，存储到 D://福布斯.csv 文件中 20 分 |
| 要求 素养 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |

| | | | |
|---|---|---|---|
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 8. H1-8，百度汽车榜爬取

（1）任务描述

基于 Python 爬虫代码从百度汽车排行榜中内容包括：汽车图片，汽车名，价格，级别，热搜指数。

网站：https://top.baidu.com/board?tab=car

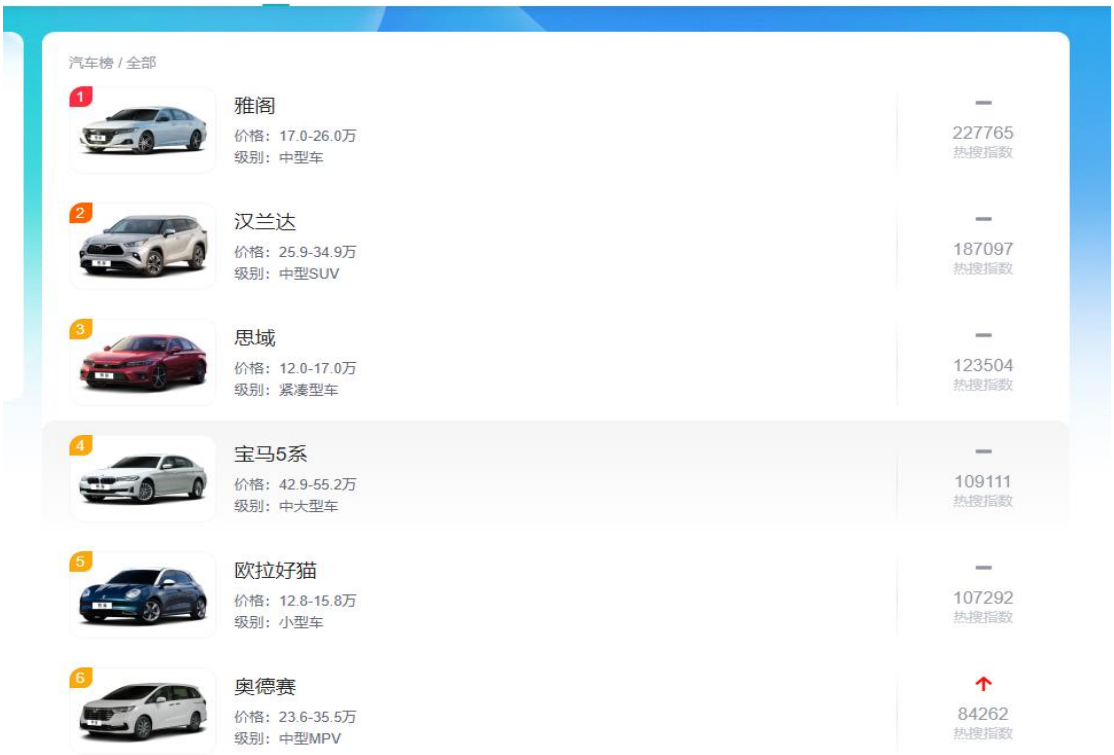


图 3.8.1 网站例图

实施步骤：

① 正确创建 scrapy 爬虫程序(10 分)。

② 正确配置 settings 文件（头部，缓存，超时等）(10 分)。

③ 正确定义 items 文件(5 分)。

④ 完成数据提取 spider 使用 xpath 或者 BS4 等完成：汽车图片，汽车名，价格，级别，热搜指数这几项数据的获取(45 分)。

⑤ 在 pipeline 中对数据汽车图片地址，汽车名，价格，级别，热搜指数进行存储，存储到 D://cars.csv 文件中，图片存储到 D://汽车 文件夹中(25 分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 10 分 | 能够正确创建 scrapy 爬虫项目 |
| | Settings 配置 | 10 分 | 1. 正确配置头部信息 5 分<br>2. 正确配置超时等基本信息 5 分 |
| | 定义 items | 5 分 | 1.根据查找数据定义对应 items 5 分 |
| | 数据提取 | 45 分 | 1. 正确创建 spider 5 分<br>2. 初始化 xpath 或 bs4 5 分<br>3. 完成汽车图片地址，汽车名，价格，级别，热搜指数这几项数据的获取 25 分<br>4. 正确转换数据为 items 10 分 |
| | 保存数据 | 25 分 | 1. 在 pipeline 中对数据汽车图片地址，汽车名，价格，级别，热搜指数进行存储，存储到 D://cars.csv 文件中 15 分<br>2. 图片存储到 D://汽车 文件夹中 10 分 |

| | | | |
|---|---|---|---|
| 素养要求 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 9. H1-9，药房网商城榜爬取

（1）任务描述

基于 Python 爬虫代码从药房网商城中西药品分类中爬取内容包括：药品图，价格，药品名，规格，批准文号，生产厂家。

网站：https://www.yaofangwang.com/catalog-1.html


图 3.9.1 网站例图

实施步骤：

① 正确创建 scrapy 爬虫程序(10 分)。

② 正确配置 settings 文件（头部，缓存，超时等）(10 分)。

③ 正确定义 items 文件(5 分)。

④ 完成数据提取 spider 使用 xpath 或者 BS4 等完成：药品图地址，价格，药品名，规格，批准文号，生产厂家 这几项数据的获取(45 分)。

⑤ 在 pipeline 中对数据价格，药品名，规格，批准文号，生产厂家进行存储，存储到 D://中西药.csv 文件中，图片存储到 D://药品 文件夹中(25 分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建项目 | 10 分 | 能够正确创建 scrapy 爬虫项目 |
| | Settings 配置 | 10 分 | 1. 正确配置头部信息 5 分<br>2. 正确配置超时等基本信息 5 分 |
| | 定义 items | 5 分 | 1.根据查找数据定义对应 items 5 分 |
| | 数据提取 | 45 分 | 1. 正确创建 spider 5 分<br>2. 初始化 xpath 或 bs4 5 分<br>3. 完成排名，名字，财富，来源，年龄，城市这几项数据的获取 25 分<br>4. 正确转换数据为 items 10 分 |
| | 保存数据 | 25 分 | 1. 在 pipeline 中对药品图，价格，药品名，规格，批准文号，生产厂家，存储到 D://中西药.csv 文件中 15 分<br>2. 图片存储到 D://药品 文件夹中 10 |

| | | | |
|---|---|---|---|
| | | | 分 |
| 素养要求 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

## 10. H1-10，当当网好评榜爬取

（1）任务描述

基于 Python 爬虫代码从当当网好评榜爬取内容包括：书本图片，书名，出版社，价格，评论数，作者。

网站：http://bang.dangdang.com/books/fivestars


图 3.10.1 网站例图

实施步骤：

① 正确创建 scrapy 爬虫程序(10 分)。

② 正确配置 settings 文件（头部，缓存，超时等）(10 分)。

③ 正确定义 items 文件(5 分)。

④ 完成数据提取 spider 使用 xpath 或者 BS4 等完成：书本图片，书名，出版社，价格，评论数，作者 这几项数据的获取(45 分)。

⑤ 在 pipeline 中对数据书名，出版社，价格，评论数，作者进行存储，存储到 D://books.csv 文件中,图片存储到 D://当当图书 文件夹中(25 分)。

作品提交：将项目文件以及数据文件保存到“所属学校_身份证_姓名_题号”命名的总文件夹中，并将文件夹打包压缩，如“娄底职业技术学院_43******_张三_题号.rar”,将压缩文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 创建项目 | 10 分 | 能够正确创建 scrapy 爬虫项目 |
| | Settings 配置 | 10 分 | 1. 正确配置头部信息 5 分<br>2. 正确配置超时等基本信息 5 分 |
| | 定义 items | 5 分 | 1.根据查找数据定义对应 items 5 分 |
| | 数据提取 | 45 分 | 1. 正确创建 spider 5 分<br>2. 初始化 xpath 或 bs4 5 分<br>3. 完成书本图片地址，书名，出版社，价格，评论数，作者这几项数据的获取 25 分<br>4. 正确转换数据为 items 10 分 |
| | 保存数据 | 25 分 | 1. 在 pipeline 中对数据书本图片地址，书名，出版社，价格，评论数，作者到 D://books.csv 文件中 15 分<br>2. 图片存储到 D://当当图书 文件夹中 10 分 |
| 素养要求 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |

| | | | |
|---|---|---|---|
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 6：网络爬虫模块实施条件

# 项目 4 Hadoop 集群部署与使用

## 1.H2-1，Hadoop 伪分布式安装与部署

（1）任务描述

Hadoop 伪分布式是一台机器，既充当 DataNode 有充当 NameNode，在单台计算机上就能模拟，方便管理和学习。搭建 Hadoop 伪分布式系统并截图记录实施步骤代码与结果。

实施步骤：

① 检查 JDK 环境(5 分)。

② 上传 Hadoop 压缩文件到/opt/soft 目录(5 分)。

③ 解压 Hadoop 压缩文件到 opt 目录，变更目录为 Hadoop(5 分)。

④ 配置 hadoop-eco.sh 环境变量，并使环境变量生效(8 分)。

⑤ 创建 Hadoop 工作目录包括（NameNode 目录，SecondaryNameNode 目录，DataNode 目录，临时数据目录）(12 分)。

⑥ 修改配置文件：修改 hadoop-env.sh 配置，yarn-env.sh 配置，core-site.xml 配置(30 分)。

⑦ 格式化 NameNode(5 分)。

⑧ 启动 HDFS(5 分)。

⑨ 启动 Yarn(5 分)。

⑩ 查看服务是否正常启动，并打开浏览器访问 Hadoop 网页(10 分)。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤从 1 到 10 截取对应操作命令与执行结果保存，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 检查 JDK 环境 | 5 分 | 命令正确，正确显示结果 |
| | 上传 Hadoop 压缩文件到/opt/soft 目录 | 5 分 | 命令正确，正确显示结果 |
| | 解压 Hadoop 压缩文件到 opt 目录，变更目录为 Hadoop | 5 分 | 命令正确，正确显示结果 |
| | 配置 hadoop-eco.sh 环境变量，并使环境变量生效 | 8 分 | 配置正确 4 分，执行生效命令 4 分<br>命令正确，正确显示结果 |
| | 创建 Hadoop 工作目录 | 12 分 | NameNode 目录，SecondaryNameNode 目录，DataNode 目录，临时数据目录每个 3 分，命令及结果正确 |
| | 修改配置文件 | 30 分 | 修改 hadoop-env.sh 配置，yarn-env.sh 配置，core-site.xml 配置每个 10 分<br>修改正确，无错误 |
| | 格式化 NameNode | 5 分 | 命令正确，正确显示结果 |
| | 启动 HDFS | 5 分 | 命令正确，服务启动完全 |
| | 启动 Yarn | 5 分 | 命令正确，服务启动完全 |
| | 查看服务是否正常启动，并打开浏览器范围 Hadoop 网页 | 10 分 | 命令正确，服务启动完全，网页能正常打开 http://ip:50070 |

| | | | |
|---|---|---|---|
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 2.H2-2，hadoop 平台架设全分布部署模块

（1）任务描述

你作为某公司运维工程师，需安装分布式 hadoop 环境。本环节需要使用 root 用户完成相关配置，具体部署要求如下:。

实施步骤:

① 解压 JDK 安装包到 " /usr/ local/src " 路径，并配置环境变量,安装 java 环境;修改环境变量配置文件并截图(10 分)。

② 创建 ssh 密钥,实现节点的无密码登录;截取主节点登录其中一个从节点的结果(10 分)。

③ 根据要求修改每台主机 host 文件，截取 " /etc/ hosts”文件内容截图(10 分)。

④ 修改每台主机 hostname 文件配置 IP 与主机名映射关系；截取“/etc/hostname”文件截图(10 分)。

⑤ 修改 Hadoop 配置 hadoop-env.sh，并截取修改内容并截图(10 分)。

⑥ 修改 Hadoop 相关文件 core-site.xml,hdfs-site.xml ,yarn-site.xml, mapred-site.xml，并初始化 Hadoop，截图初始化结果(20 分)。

⑦ 启动 Hadoop，使用相关命令查看所有节点 Hadoop 进程并截图(20 分)。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，按步骤截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | Jdk 解压与配置 | 10 分 | 1. 正确安装 jdk 5 分<br>2. 正确配置 java 环境变量 5 分 |
| | Ssh 安装及访问 | 10 分 | 1. 正确安装与配置 Ssh 5 分<br>2. 完成 ssh 连接 5 分 |
| | Host 配置 | 10 分 | 1. 正确配置 Host 文件 10 分 |
| | IP 与主机名映射关系 | 10 分 | 1. 正确配置主机映射关系 10 分 |
| | 配置 hadoop-env.sh | 10 分 | 1. 正确配置 hadoop-env.sh 10 分 |
| | Hadoop 其他配置文件 | 20 分 | 1. core-site.xml,hdfs-site.xml ,yarn-site.xml, mapred-site.xml 每个配置文件 5 分 |
| | Hadoop 初始化截图 | 20 分 | 1. 正确格式化 namenode 5 分<br>2. 正确执行启动 5 分<br>3. 运行结果正确 6 个服务开启正常 10 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 3.H2-3，hadoop 平台架设 Hbase 组件部署模块

（1）任务描述

你作为某公司运维工程师，需在已安装 hadoop 环境下部署 hbase。本环节需要使用 root 用户完成相关配置，具体部署要求如下:。

实施步骤:

① 解压 Hbase 安装包到“/usr/local/src”路径，并修改解压后文件夹名为 hbase，截图并保存结果。

② 设置 Hbase 环境变量，并使环境变量只对当前 root 用户生效，截图并保存结果。

③ 修改 Hbase 相应配置文件，截图并保存结果。

④ 把 Hadoop 的相应文件放到 hbase/conf 下，截图并保存结果。

⑤ 启动 Hbase 并保存命令输出结果，截图并保存结果。

⑥ 创建 Hbase 数据库表，截图并保存结果。

| 表 t_temp | |
|---|---|
| 字段名 | 字段说明 |
| empno | 编号 |
| ename | 姓名 |
| job | 工作 |
| sal | 工资 |
| deptno | 部门 |
| | |

⑦ 将给定数据写入数据库表中，截图并保存结果。

| empno | ename | job | sal | deptno |
|---|---|---|---|---|
| 701 | TCS | Research Scientist | 1009 | 20 |
| 1012 | Accenture | Research Scientist | 1051 | 20 |
| 1056 | Cognizant | Sales Representative | 1052 | 10 |
| 1876 | ICICI Bank | Sales Representative | 1081 | 10 |
| 1928 | HDFC Bank | Sales Representative | 1091 | 10 |
| 243 | Wipro | Laboratory Technician | 1102 | 20 |

| | | | | |
|---|---|---|---|---|
| 1273 | Infosys | Sales Representative | 1118 | 10 |
| 1974 | Capgemini | Laboratory Technician | 1129 | 20 |
| 411 | Tech Mahindra | Sales Representative | 1200 | 10 |

⑧ 查看 Hbase 版本信息，截图并保存结果。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，按步骤截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | Hbase 解压 | 10 分 | 没有正确解压到指定位置扣 10 分 |
| | 设置环境变量 | 10 分 | 环境变量不正确扣 10 分 |
| | Hbase 配置文件 | 10 分 | 环境变量每少一个扣 3 分 |
| | Hadoop 相关包 | 10 分 | 复制包每少一个扣 2 分 |
| | 启动 hbase | 10 分 | 不能正常启动扣 10 分 |
| | Hbase 数据库 | 10 分 | 未按要求建立表结构扣 4 分 |
| | 导入数据 | 20 分 | 数据导入不成功每条数据扣 2 分 |
| | Hbase 版本信息 | 10 分 | 无法查看版本信息扣 10 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 4.H2-4，hadoop 平台架设 Hive 组件部署模块

（1）任务描述

本环节需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体部署要求如下:。

实施步骤：

① 解压 Hive 安装包到“/usr/local/src”路径，并使用相关命令，修改解压后文件夹名为 Hive，进入 Hive 文件夹，并将查看内容截图(10 分)。

② 设 置 Hive 环境变量 （ HIVE_HOME=/usr/local/src/hive PATH=$PATH:$HIVE_HOME/bin），并使环境变量只对当前用户生效(10 分)。

③ 新建并配置 hive-site.xml 文件，实现“Hive 元存储”的存储位置为 MySQL 数据库(20 分)。

④ 初始化 Hive 元数据（将 MySQL 数据库 JDBC 驱动拷贝到 Hive 安装目录的 lib 下），初始化结果截图(10 分)。

⑤ 启动 Hive，检查是否安装成功，截图保存结果(10 分)。

⑥ 按指定要求创建 Hive 内部表和外部表，截图保存结果(20 分)。

| 外部表 user | |
|---|---|
| 字段名 | 字段说明 |
| userid | 编号 |
| username | 姓名 |
| nikename | 昵称 |
| userage | 年龄 |

| 内部表 inneruser | |
|---|---|
| 字段名 | 字段说明 |
| inneruserid | 编号 |
| innerusername | 姓名 |
| innernikename | 昵称 |
| inneruserage | 年龄 |

⑦ 实现内外部表转换，截图保存结果(10 分)。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，按步骤

截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | Hive 解压 | 10 分 | 未成功解压到指定位置扣 10 分 |
| | 设置环境变量 | 10 分 | 环境变量设置不正确每点扣 3 分 |
| | Hive 配置文件 | 20 分 | 正确配置 hive-site.xml 20 分 |
| | 初始化 hive | 10 分 | 未拷贝驱动扣 10 分 |
| | 启动 hive | 10 分 | 启动不成功扣 10 分 |
| | 创建表结构 | 20 分 | 内部表 10 分,外部表 10 分 |
| | 实现内外表转换 | 10 分 | 未按要求实现转换扣 10-20 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 5.H2-5，hadoop 平台架设 Flume 模块

（1）任务描述

你作为某公司运维工程师，需在已安装 hadoop 环境下部署 Flume。本环节需要使用 root 用户完成相关配置，具体部署要求如下:。

实施步骤:

① 解压 Flume 安装包到“/usr/local/src”路径，并修改解压后文件夹名 flume (10 分)。

② 设置 Flume 环境变量，并使环境变量只对当前用户生效并截图 (20 分)。

③ 修改 Flume 相应文件 `flume-env.sh` 并截图 (20 分)。

④ 启动 Flume 并检测 Flume 版本并截图(10 分)。

⑤ 完成 Flume 的 agent,source,channel,sink 配置并截图(30 分)。

⑥ 通过 Flume 将 weblog.log 中数据传输到 HDFS 中，截图并保存结果(10 分)。

作品提交:创建“所属学校_身份证_姓名_题号”命名的 word 文档，按步骤截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | Flume 解压 | 10 分 | 1. 正确解压 Flume 5 分<br>2. 正确修改 Flume 文件夹名称 5 分 |
| | 设置环境变量 | 20 分 | 1. 正确配置环境变量 15 分<br>2. 正确设置环境变量生效 5 分 |
| | Flume 配置文件 | 20 分 | 1. 正确配置 `flume-env.sh` |

| | | | |
|---|---|---|---|
| | 启动 Flume | 10 分 | 1. 正确启动 Flume 5 分<br>2. 查看 Flume 版本且正常显示 5 分 |
| | 配置 Flume 环境 | 20 分 | 1. 正确设置 agent 5 分<br>2. 正确设置 source 10 分<br>3. 正确设置 channel 5 分<br>4. 正确设置 sink 10 分 |
| | Flume 数据传输 | 10 分 | 1. 正确启动 flume 并上传文件 10 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 6.H2-6，hadoop 平台架设 kafka 组件部署模块

（1）任务描述

你作为某公司运维工程师，需在已安装 hadoop 环境下部署 Kafka。本环节需要使用 root 用户完成相关配置，具体部署要求如下:。

实施步骤：

① 启动 Zookeeper 并截图保存结果(5 分)。

② 解压 Kafka 安装包到“/usr/local/src”路径，并修改解压后文件夹名为 kafka，截图并保存结果(10 分)。

③ 设置 Kafka 环境变量，并使环境变量只对当前 root 用户生效，截图并保存结果(10 分)。

④ 修改 Kafka 相应文件，截图并保存结果(10 分)。

⑤ 启动 Kafka 并保存命令输出结果，截图并保存结果(5 分)。

⑥ 创建指定 topic，并截图并保存结果(5 分)。

⑦ 查看所有的 topic 信息，并截图并保存结果(5 分)。

⑧ 启动指定生产者（producer)，并截图并保存结果(10 分)。

⑨ 启动消费者（consumer），并截图并保存结果(10 分)。

⑩ 测试生产者（producer），并截图并保存结果(10 分)。

⑪ 测试消费者（consumer），并截图并保存结果(10 分)。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，按步骤截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 启动 zookeeper | 5 分 | 1. 正确启动 zookeeper 5 分 |
| | 解压 kafka | 10 分 | 1. 正确解压 kafka 5 分<br>2. 正确修改文件夹 5 分 |
| | 设置环境变量 | 10 分 | 1. 正确设置环境变量 10 分 |
| | 修改文件并启动 kafka | 15 分 | 1. 正确修改 server.properties 10 分<br>2. 正确启动 kafka 5 分 |
| | 创建和查看 topic | 10 分 | 1. 正确创建 topic 5 分<br>2. 正确查看 topic 5 分 |
| | 启动生产者 | 10 分 | 1. 正确启动生产者 10 分 |
| | 启动消费者 | 10 分 | 1. 正确启动消费者 10 分 |
| | 测试生产者 | 10 分 | 1. 成功测试生产者 10 分 |
| | 测试消费者 | 10 分 | 1. 成功测试消费者 10 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 7.H2-7，使用 Hadoop 进行词频统计

（1）任务描述

词频统计是大数据操作的最常见的操作，经常用于用户的费用统计，销售统计等操作，现有一组数据文件，文件中存在很多单词，结构形式如下所示，现要求使用 Hadoop 环境将单词进行词频统计，统计出每个单词的数量。

实施步骤：

任务一：词频统计准备（36 分）

① 使用 linux 命令在当前用户主目录下创建 input 文件夹（10 分）。

② 使用 file1.txt 和 file2.txt，上传到 input 文件夹中（6 分）。

③ 启动 Hadoop 的 hdfs，使用 hdfs 命令在 hdfs 的根目录上创建文件夹 test_input（10 分）。

④ 使用 hdfs 的 shell 命令将 file1.txt,file2.txt 上传到 hdfs 的文件夹 test_input 中（10 分）。

任务二：编写词频统计代码（40 分）

① 编写词频统计 Mapper 代码（15 分）。

② 编写词频统计 Reduce 代码（15 分）。

③ 编写程序入口 Main 部分代码（5 分）。

任务三：启动 hadoop 并运行得到结果（24 分）

① 将编写的代码打成 test.jar 包（8 分）。

② 启动 hadoop，并确认 hadoop 已启动.（4 分）。

③ 在 hadoop 上通过命令运行 test.jar 包，并得到执行结果（12 分）。

作品提交：创建“所属学校_身份证_姓名_题号”命名的文件夹，“词频统计准备”过程保存与结果.doc 文件中，并把对应 java 代码保存到文件夹中，压缩文件夹上传提交到对应位置。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 词频统计准备 | 35 分 | 1、创建 input 文件夹（10）<br>2、上传文件到对应位置（6）<br>3、在 hdfs 上创建文件夹（5）<br>4、文件上传（10） |
| | 编写词频统计代码 | 35 分 | 1、编写 Mapper 代码（15）<br>2、编写 Reduce 代码（15）<br>3、编写程序入口 Main 部分代码（5） |
| | 启动 hadoop 并运行得到结果 | 20 分 | 1、代码打包（6）<br>2、启动 hadoop（4）<br>3、代码运行并输出结果（10） |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 8.H2-8，朝阳医院销售数据清洗

（1）任务描述

本题假设以朝阳医院 2018 年销售数据为例，目的是了解朝阳医院在 2018 年里的销售情况，通过对朝阳区医院的药品销售数据的分析，了解朝阳医院的患者的月均消费次数，月均消费金额、客单价以及消费趋势、需求量前几位的药品等。

本任务为分析前数据准备：处理数据表中空值数据，数据类型异常值

实施步骤：

① 检查 Hadoop 环境（启动查看对应服务），JDK 环境。

② 观察“朝阳医院 2018 年销售数据.xlsx”数据，寻找可能存储缺失值，类型异常的信息，确定各列所代表含义。

③ 编写实例类用来记录数据中的字段。

④ 编写自定义 Mapper 类对“购买时间”,“社保卡号”空值进行删除处；“销量”,“应收金额”,“实收金额”类型统一为小数类型，若为负数则删除该条数据。

⑤ 自定义 Reduce 类处理信息。

⑥ 自定义 Driver 类处理输入，输出，输出文件夹为“朝阳医院数据结果”。

⑦ 正确上传 jar 并运行完成数据处理得到结果。

作品提交：创建“所属学校_身份证_姓名_题号”命名的文件夹，包含 java 程序，数据处理结果文件,将文件压缩并按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 检查 JDK 环境 | 5 分 | 命令正确，正确显示结果 |
| | 检查 Hadoop 环境 | 5 分 | 命令正确，正确显示结果 |
| | 实例类 | 10 分 | 根据表格字段编写对应实例类 |
| | Mapper 类 | 30 分 | 1、正确创建自定义 Mapper 类 5 分<br>2、正确处理购买时间，社保卡号空值信息 10 分<br>3、正确“销量”，“应收金额”，“实收金额”类型统一为小数类型，若为负数则删除该条数据 15 分 |
| | Reduce 类 | 15 分 | 正确创建自定义 Reduce 类并编写正确处理逻辑 |

| | | | |
|---|---|---|---|
| | Driver 类 | 15 分 | 正确自定义 Driver 类并实现文件的输入与输出 15 分 |
| | 运行结果 | 10 分 | 正确提交 jar 并运行得到对应结果 |
| 素质要求 | 代码书写规范 | 3 分 | 代码格式规范，缩进合理 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 9.H2-9，使用 Hadoop 实现求平均成绩

（1）任务描述

数据的处理有多种多样的表达，求取平均值是最常见的数据处理之一，对输入文件中数据进行计算学生平均成绩。输入文件中的每行内容均为一个学生的姓名和他相应的成绩，如果有多门学科，则每门学科为一个文件。要求在输出中每行有两个间隔的数据，其中，第一个代表学生的姓名，第二个代表其平均成绩。

实施步骤：

任务一：大数据准备（36 分）

① 使用 linux 命令在当前用户主目录下创建 input 文件夹（10 分）。

② 使用 vim 编辑在 input 文件夹下创建 math.txt 和 chinese.txt，按照提示输入内容并保存（6 分）。

③ 启动 Hadoop 的 hdfs，使用 hdfs 命令在 hdfs 的根目录上创建文件夹 score_input（10 分）。

④ 使用 hdfs 的 shell 命令将 math.txt,chinese.txt 上传到 hdfs 的文件夹 score_input 中（10 分）。

任务二：编写平均值统计代码（30 分）

① 编写平均值 AvgMapper 代码（15 分）。

② 编写平均值 AvgReduce 代码（10 分）。

③ 编写程序入口 AvgMain 部分代码（5 分）。

任务三：启动 hadoop 并运行得到结果（24 分）

① 将编写的代码打成 avg.jar 包（8 分）。

② 启动 hadoop，并确认 hadoop 已启动（4 分）。

③ 在 hadoop 上通过命令运行 avg.jar 包，通过命令参数指定输入和输出目录，并得到执行结果（12 分）。

作品提交：创建“所属学校_身份证_姓名_题号”命名的文件夹，“大数据准备”过程保存与 结果.doc 文件中，并把对应 java 代码保存到文件夹中，压缩文件夹上传提交到对应位置。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 大数据准备 | 36 分 | 1、创建 input 文件夹（10）<br>2、上传文件到对应位置（6）<br>3、在 hdfs 上创建文件夹（6）<br>4、文件上传（10） |
| | 编写平均值统计代码 | 30 分 | 1、编写 AvgMapper 代码（15）<br>2、编写 AvgReduce 代码（10）<br>3、编写程序入口 AvgMain 部分代码（5） |
| | 启动 hadoop 并运行得到结果 | 24 分 | 1、代码打包（8）<br>2、启动 hadoop（4） |

| | | | |
|---|---|---|---|
| | | | 3、代码运行并输出结果（12） |
| 素质要求 | 代码书写规范 | 3 分 | 代码格式规范，缩进合理 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 结果提交 | 5 分 | 截图清晰，项目打包正确，结果完备 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 10.H2-10，使用 Hadoop 求销售额排名前 5 位的销售纪录

（1）任务描述

销售数据是大数据分析处理当中最常见的数据，尤其是对数据做 TopN 的操作更是数据处理中必备的操作数据，先要求对销售记录文件进行排序，得到销售记录最高的 5 条记录，销售记录文件格式如下：

Orderid（订单号）Userid（用户 id） Payment（销售额） Productid（产品 id)。

实施步骤：

任务一：大数据准备（36 分）

① 使用 linux 命令在当前用户主目录下创建 input 文件夹（10 分)。

② 向 input 文件夹下上传 sales1.txt 和 sales2.txt 文件（6 分)。

③ 启动 Hadoop 的 hdfs，使用 hdfs 命令在 hdfs 的根目录上创建文件 sort_input 文件夹（截图）(10 分)。

④ 使用 hdfs 的 shell 命令将 sales1.txt,sales2.txt 上传到 hdfs 的文件

夹 sort_input 文件夹中（截图）(10 分)。

任务二：编写平均值统计代码（30 分）

① 编写排序 InvertedMapper 代码（15 分)。

② 编写排序 InvertedReduce 代码（10 分)。

③ 编写程序入口 InvertedMain 部分代码。(5 分)。

任务三：启动 hadoop 并运行得到结果（24 分）

① 将编写的代码打成 inverted.jar 包（8 分)。

② 启动 hadoop，并确认 hadoop 已启动（4 分)。

③ 在 hadoop 上通过命令运行 inverted.jar 包，通过命令参数指定输入和输出目录，并得到执行结果（12 分)。

作品提交：创建“所属学校_身份证_姓名_题号”命名的文件夹，“大数据准备”过程保存与 结果.doc 文件中，并把对应 java 代码保存到文件夹中，压缩文件夹上传提交到对应位置。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 大数据准备 | 36 分 | 1、创建 input 文件夹（10）<br>2、上传文件到对应位置（6）<br>3、在 hdfs 上创建文件夹（6）<br>4、文件上传（10） |
| | 编写平均值统计代码 | 30 分 | 1、编写 InvertedMapper 代码（15）<br>2、编写 InvertedReduce 代码（10）<br>3、编写程序入口 InvertedMain 部分代码（5） |
| | 启动 hadoop 并运行得到结果 | 24 分 | 1、代码打包（8）<br>2、启动 hadoop（4）<br>3、代码运行并输出结果（12） |

| | | | |
|---|---|---|---|
| 素质要求 | 代码书写规范 | 3 分 | 代码格式规范，缩进合理 |
| | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | 结果提交 | 5 分 | 截图清晰，项目打包正确，结果完备 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

# 项目 5 数据仓库 Hive 部署与使用

## 1.H3-1，奇书网脏数据处理

（1）任务描述

使用 Hive 对奇书网数据进行数据导入以及对应查询操作。

实施步骤：

① 使用 Hive 创建数据库 qishu。

② 使用 Hive 创建数据表：book,book_type,book_info 要求：按照数据文件的陪陪字段要求创建表字段，字段类型合理。

③ 将 qishu_book.csv 文件数据导入 hive 表 book。

④ 将 qishu_type.csv 文件数据导入 hive 表 book_type 中。

⑤ 将 qishu_info.csv 文件数据导入表 book_info 。

⑥ 按照点击量降序查询点击率前 10 的小说。

⑦ 查询每个类型的小说的点击量总量，按降序输出。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤从 1 到 5 截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 创建数据库 | 5 分 | 使用 hive 正确创建数据库 qishu,命令和结果正确 5 分 |
| | 创建数据表 | 20 分 | 使用 hive 创建表：book,book_type,book_info 三张表，表字段符合数据要求；命令和结果正确；<br>表 book 5 分<br>表 book_type 5 分<br>表 book_info 10 分 |
| | 导入数据 | 25 分 | 使用 hive 将对应 csv 文件数据正确导入数据表；命令和结果正确。<br>导入 book_type 表 5 分<br>导入 book 表 10 分<br>导入 book_info 表 10 分 |
| | 查询数据 | 40 分 | 1. 按照点击量降序查询点击率前 10 的小说，命令和结果正确 20 分，命令基本正确得 5 分<br>2. 查询每个类型的小说的点击量总量，按降序输出，命令和结果正确 20 分，命令基本正确得 5 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 2.H3-2，奇书网特殊数据处理

（1）任务描述

使用 Hive 对奇书网数据进行数据导入以及对应查询操作。

实施步骤：

① 使用 Hive 创建数据库 qishu。

② 使用 Hive 创建数据表：book,book_type,book_info 要求：按照数据文件的陪陪字段要求创建表字段，字段类型合理。

③ 将 qishu_book.csv 文件数据导入 hive 表 book。

④ 将 qishu_type.csv 文件数据导入 hive 表 book_type 中。

⑤ 将 qishu_info.csv 文件数据导入表 book_info 。

⑥ 查询“武侠仙侠”类型小说的点击量排行，按照升序排序。

⑦ 查询每个类型的小说总量，按照总量降序排序。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤从 1 到 5 截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建数据库 | 5 分 | 使用 hive 正确创建数据库 qishu,命令和结果正确 5 分 |
| | 创建数据表 | 20 分 | 使用 hive 创建表：book,book_type,book_info 三张表，表字段符合数据要求；命令和结果正确；<br>表 book 5 分<br>表 book_type 5 分<br>表 book_info 10 分 |
| | 导入数据 | 25 分 | 使用 hive 将对应 csv 文件数据正确导入数据表；命令和结果正确。 |

| | | | |
|---|---|---|---|
| | | | 导入 book_type 表 5 分<br>导入 book 表 10 分<br>导入 book_info 表 10 分 |
| | 查询数据 | 40 分 | 1. 查询“武侠仙侠”类型小说的点击量排行，按照升序排序，命令和结果正确 20 分，命令基本正确得 5 分<br>2. 查询每个类型的小说总量，按照总量降序排序，命令和结果正确 20 分，命令基本正确得 5 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 3.H3-3，员工信息处理

（1）任务描述

现有一份职工数据 emp.csv,记录了公司职工的姓名、职工编号等信息，部分数据如下 表。(注：emp.csv 的数据分割符为“,”mpno 为职工编号，ename 为职工姓名，job 为工作职 位，sal 为薪资，deptn。为员工所在部门编号]

| empno | ename | job | sal | deptno |
|---|---|---|---|---|
| 7369 | SXQTH | CLERK | SOO | 30 |
| 7499 | ALLEN | SALESMAN | 1600 | 20 |
| 7566 | JONES | MANAGER | 2975 | 30 |
| 7654 | MARTIN | SALESMAN | 1250 | 30 |
| 769S | BLAKE | MANAGER | 2850 | 10 |
| 7839 | KING | PRESIDENT | 5000 | 20 |

观察该表字段和 Hive 表中的具体数据，完成以下任务

实施步骤：

① 观察数据，在 Hive 中创建对应字段数据的表（以“，"分隔)，命名为 emp,要求提交创建表代码。

② 将Linux本池路径/course/Hive/data/Temp.csv数据导入所创建的emp数据表 中，要求提交号入数据的代码，并駁取查看 emp 表数据的结果截图使用 load data local inpath *** overwrite into table 函数导入数据。

③ 将 emp_in 表按照职工编号升序、薪资降序方式送行排列。求提交对数据送行排序的代码，关截取数据排序的结果裁图，(使用 sort by 方法)。

④ 查询薪资大于 3000 的工作职位（字段 job）的数据。要求提交查询数据的代码，并截取查询数据的结果截图。

⑤ 查询 10 号部门中，薪资大于 1500 的职工姓名，要求提交查询数据的代码并截取查询数据截图。

⑥ 统计每个部门的员工数。要求提交查询数据的代码，并截取查询数据的结果截图。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步

骤从 1 到 5 截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 创建表 | 15 分 | 成功创建 15 分，类型错误每个扣 1 分 |
| | 查询语句及结果 | 15 分 | 结果截图正确满分，代码出错依情况扣 10 到 5 分，使用 sort by 正确得 5 分，定义为降序排序 |
| | 查询语句及结果 | 20 分 | 结果截图正确满分，代码出错依情况扣 10 到 5 分，使用 group by 函数分组的 5 分 |
| | 查询语句及结果 | 20 分 | 结果截图正确满分，使用 where 条件进行是筛选约束各的 8 分，未使用扣除。结果不正确语句基本正确扣 4 分。 |
| | 统计及结果 | 20 分 | 结果截图正确得满分，使用 group by 函数进行分组得 8 分，成功统计各组人员得 9 分，结果错误扣 3 分 |
| | 文档规范 | 10 分 | 截图完整，排版规范合理，代码命名规范，变量名命名规范。 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 4.H3-4，学生数据处理

（1）任务描述

现有一份学生数据 student.csv 该数据具体字段如下。

| 字段名称 | 字段说明 | 字段类型 | 数据示例 |
|---|---|---|---|
| sname | 姓名 | string | SMITH. ALLEN. ... |
| sno | 学号 | string | sOl, s02, s03. ... |
| sex | 性别 | int | 0. 1 (0 为男，1 为女) |
| age | 年龄 | int | 16. 19. 18.... |

观察该表字段和 Hive 表中的具体数据，任务如下。

实施步骤：

① 进入 hive 命令行创建数据表；或根据字段信息编写脚本文件，创建对应字段名及 字段类型的 Hive 表〔需设置表的分割字符为", "），并将数据表命名为 student，要求提交创建表的语句。

② 将 Linux 本地路径/course/Hive/data/Tˆ)/student.csv 文件导入到职创建的 Hive 表中，要求提交号入数据的语句，关查看数据表，获取导入数据后数据表的内容截 图：（使用 load data local inpath *** overwrite into table 函数逐行数据导入）。

③ 查询性别为男，且年龄大于 18 的数据。要求提交查询数据的代码，并裁取查询数 据的结果截图：（使用 where 方法）。

④ 查询性别为女的数据记录数：要求提交查询数据的代码，获取查询数据的结果截图（使用 count 函数和 where 方法）。

⑤ 查询性别为另 M 年龄小于 18 的数据记录数，要求提交查询数据的代码，芦駁取查 询数据的结果截图（使用 count 函数和 where 方法）。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤从 1 到 5 截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 创建表 | 15 分 | 成功创建 15 分，类型错误每个扣 1 分 |
| | 查询语句及结果 | 15 分 | 结果截图正确满分，代码出错依情况扣 10 到 5 分，使用 sort by 正确得 5 分，定义为降序排序 |
| | 查询语句及结果 | 20 分 | 结果截图正确满分，代码出错依情况扣 10 到 5 分，使用 group by 函数分组的 5 分 |
| | 查询语句及结果 | 20 分 | 结果截图正确满分，使用 where 条件进行是筛选约束各的 8 分，未使用扣除。结果不正确语句基本正确扣 4 分。 |
| | 统计及结果 | 20 分 | 结果截图正确得满分，使用 group by 函数进行分组得 8 分，成功统计各组人员得 9 分，结果错误扣 3 分 |
| | 文档规范 | 10 分 | 截图完整，排版规范合理，代码命名规范，变量名命名规范。 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

## 5.H3-5，基站信息处理

（1）任务描述

现有一份数据 business_circle.csv,数据为基于基站，对各商圈内的人流进行统计的数据，具体字段信息如下表：

| 字段名 | 字段说明 | 字段类型 | 数据示例 |
|---|---|---|---|
| ID | 基站编号 | Int | 36902, 36903, 36904… |
| Time_workday | 工作日上班时间人均停留时间 | Int | 78,144,95,69,190... |
| Time_midnight | 凌晨人购停留时间 | Int | 521,600,457... |
| Time_weekend | 周末人均停留时间 | Int | 601521,468... |
| Time_day | 日均人流量 | Int | 2S63^245J054... |

观察该表字段和 Hive 表中的具体数据’完成以下任务，

① 上表字段信息，在 Hive 中创建对应字段名及字段类型的表.命名为 business，要求提交创建表的代码。

② 将 Linux 本地路径/course/Hive/data/下的 business_circle.csv 数据导入到 business 表，并查看导入数据后的数据表的前 10 行数据要求提交导入数据的代码和查询数据的结果截图：（使用 load data local inpath *** overwrite into table 函数逐行数据导入。

③ 查询凌晨人均停留旳间大于 500 的数据记录。要求提交查询数据的代码，并截取查询数据的结果截图 。

④ 查询日均人流星大于 1500 的数据记录，要求提交查询数据的代码，并截取查询数据的結果截图。

⑤ 查询人流最大及人流是最小对应的数据记录：要求提交查询数据的代码，并截取查询数据的结果截图。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步

骤从 1 到 5 截取对应操作命令与执行结果，将文件按要求上传。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建表 | 15 分 | 成功创建 15 分，类型错误每个扣 1 分 |
| | 查询语句及结果 | 15 分 | 数据导入成功并成功查询得 15 分，数据导入代码 10 分，数据查询代码 5 分，查询结果显示不全 3 分 |
| | 查询语句及结果 | 15 分 | 结果截图正确满分，代码出错依情况扣 10 到 5 分，条件函数使用正确得 5 分 |
| | 查询语句及结果 | 20 分 | 结果截图正确满分，使用 where 条件进行是筛选约束各的 8 分，未使用扣除。结果不正确语句基本正确扣 4 分。 |
| | 统计及结果 | 25 分 | 结果截图正确得满分，使用 max,min,where 函数，每个得分 5 分，结果错误扣 5 分 |
| | 文档规范 | 10 分 | 截图完整，排版规范合理，代码命名规范，变量名命名规范。 |

（4）实施条件

见附录 7：Hadoop 平台与组件模块实施条件

# 项目 6 Flink 的部署与使用

## 1. H4-1，Flink 流处理

（1）任务描述

通过 Socket 手工实时产生一些单词，使用 FLink 实时接收数据，对指定时间窗口内（如 2s）的数据进行聚合统计，并且把时间窗口内计算的结果打印。

实施步骤：

① 搭建 flink 项目。

② 获取需要的端口号。

③ 获取 Flink 的运行环境。

④ 连接 Socket 获取输入的数据。

⑤ 把数据打印到控制台并设置并行度。

⑥ 调用实现并截图结果。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 1、搭建 flink 项目 | 10 分 | 正确配置 Maven，并搭建 Flink 项目 |
| | 2、获取需要的端口号。 | 10 分 | 代码正确 |
| | 3、获取 Flink 的运行环境 | 15 分 | 代码正确 |
| | 4、连接 Socket 获取输入的数据 | 40 分 | 正确使用 DataStreamSource 5 分<br>正确使用 flatMap 10 分<br>正确使用 DataStream 15 分 |

| | | | |
|---|---|---|---|
| | | | 正确使用 timeWindow 10 分 |
| | 5、把数据打印到控制台并设置并行度 | 10 分 | 代码正确 |
| | 6、调用实现 | 5 分 | 正确调用 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 8：数据分析模块实施条件

## 2. H4-2，Flink 批处理单词统计

（2）任务描述

统计 word.txt 文件中的单词出现的总次数，并且把结果存储到文件中。

实施步骤：

① 搭建 Flink 项目。

② 获取运行环境。

③ 获取文件中内容。

④ 转换为 DataSet 并处理。

⑤ 使用 FlatMap 完成统计。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 别 | | | |
| 技能要求 | 1.搭建 Flink 项目 | 10 分 | 正确搭建 |
| | 2.获取运行环境 | 15 分 | 代码正确 |
| | 3.获取文件中内容 | 15 分 | 代码正确 |
| | 4.转换为 DataSet 并处理 | 10 分 | 4、正确使用 flatmap 10<br>5、正确使用 groupby 10<br>6、正确使用保存处理 10 |
| | 5.使用 FlatMap 完成统计 | 20 分 | 1、正确实现 FlatMapFunction 接口（10 分）<br>2、正确完成单词统计（10 分） |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 8：数据分析模块实施条件

# 项目 7 Spark 的部署与使用

## 1.H5-1，使用 Spark 进行数据去重和处理

（1）任务描述

有部门数据文件 dept.txt 和员工数据文件 employee.txt 两个数据文件，部门数据文件的格式为 dept.txt(deptid,deptname)，员工数据文件 employee.txt 结构为 employee.txt(eid,name,score(绩效),salary,deptid)， 部门数据文件 dept.txt 的数据如下：

1,人事部

2,销售部

3,财务部

员工数据文件 employee.txt 的数据如下：

1,张三,80,3500,1

2,李白,90,4500,1

3,狄仁杰,88,5500,2

4,,92,6500,2

5,小乔,69,7500,2

6,貂蝉,86,9500,2

7,,87,5500,3

8,后羿,89,4500,3

9,蔡文姬,80,3800,3

实施步骤：

任务一：RDD 数据创建及清洗（45 分）

① 读取 dept.txt 和 employee.txt 创建两个 RDD（15 分）。

② 对 employee 的 RDD 进行去空处理，删除 name 字段为空的记录（15 分）。

③ 对创建的两个 RDD 进行数据去重（15 分）。

任务二：对 RDD 进行统计处理（45 分）

① 对两个 RDD 根据 deptid 字段进行内连接的操作（10 分）。

② 统计并打印出每个部门薪资的总额（15 分）。

③ 打印出每个部门绩效平均分（20 分）。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 创建 RDD; | 15 分 | 数据文件的 RDD 正确创建 |
| | 对 RDD 数据进行去空 | 15 分 | 正确去除 name 为空的字段 |
| | 对 RDD 数据进行去重 | 15 分 | 去除重复数据 |
| | RDD 连接操作 | 10 分 | 部门 RDD 和员工 RDD 连接操作 |
| | 计算部门薪资总额 | 15 分 | 正确进行 Map 操作，并生成输出键值对 |
| | 计算部门绩效平均值 | 20 分 | 根据 Mapper 输入得到正确的平均值结果 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 8：数据分析模块实施条件

## 2.H5-2，使用 Spark 进行日志数据分析

（1）任务描述

在一个电商平台中，每天都会产生大量的日志数据，日志数据反映了用户对平台的访问的操作细则，电商平台需要对后台的日志数据进行分析，区别统计 Get 和 Post 的 url 的访问量，要求分析结果格式为 访问方式，url，访问量日志文件见 webdata.log。

实施步骤：

任务一：创建数据文件（45 分）

① 将目标 webdata.log 导入到对应位置（10 分）。

② 正确创建 SparkContext 对象，设置部署模式为本地模式（20 分）。

③ 正确读取 data.txt 形成 RDD（15 分）。

任务二：RDD 算子操作（45 分）

① 对数据进行清洗，去掉来自其他机器的访问数据（10 分）。

② 对数据进行数据截取操作，获取数据相关的列（15 分）。

③ 对数据进行转换，行动，统计操作，得到最终结果（20 分）。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 导入数据文件 | 10 分 | 正确导入数据文件 |
| | 初始化 Spark 框架 | 20 分 | 正确创建 SparkContext 对象 |
| | 创建 RDD | 15 分 | 读取数据文件创建 RDD |

| | | | |
|---|---|---|---|
| | 数据清洗 | 10 分 | 去除规定的数据 |
| | 数据截取 | 15 分 | 截取需要的数据列 |
| | 数据分析处理 | 20 分 | 数据转换，统计，行动操作 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 8：数据分析模块实施条件

## 3.H5-3，使用 Spark 操作 MySQL 数据库数据

（1）任务描述

在大数据平台的数据处理，通常都会对数据的处理结果进行保存，通常的数据保存可以选择保存在 HDFS 或者 MySQL 数据库当中，也通常会从 MySQL 数据库中读取数据并进行处理，所以 Spark 需要对 MySQL 进行读写操作，本次任务是从 MySQL 数据库中读取数据，进行处理，并将结果写回到 MySQL 数据库。

该任务场景为业务系统的销售记录表（sales），需要统计出不同地区的销售额并进行升序排序，并将数据插入到统计结果表（result）数据库的表结构如下：

| 字段名称 | 字段说明 |
|---|---|
| Saleid | 销售 id |
| Saletime | 售卖时间 |
| SaleUser | 售货员 |
| SaleAre | 地区 |
| SaleCount | 售卖件数 |

实施步骤：

任务一：创建数据文件（45 分）

① 将目标 webdata.log 导入到对应位置（10 分）。

② 正确创建 SparkContext 对象，设置部署模式为本地模式（20 分）。

③ 正确读取 data.txt 形成 RDD（15 分）。

任务二：RDD 算子操作（45 分）

① 对数据进行清洗，去掉来自 8.8.8.8 的访问数据（10 分）。

② 对数据进行数据截取操作，获取数据相关的列（15 分）。

③ 对数据进行转换，行动，统计操作，得到最终结果（20 分）。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|
| 技能要求 | 导入数据文件 | 10 分 | 正确导入数据文件 |
| | 初始化 Spark 框架 | 20 分 | 正确创建 SparkContext 对象 |
| | 创建 RDD | 15 分 | 读取数据文件创建 RDD |
| | 数据清洗 | 10 分 | 去除规定的数据 |
| | 数据截取 | 15 分 | 截取需要的数据列 |
| | 数据分析处理 | 20 分 | 数据转换，统计，行动操作 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 |

（4）实施条件

见附录 8：数据分析模块实施条件

# 模块 3．拓展岗位技能模块

## 项目 8 Python 数据可视化

### 1.Z1-1，超市销售数据可视化与分析

（1）任务描述

supermarket_sales.xlsx 是某超市 2015 年 1 月 1 日至 4 月 30 日的经营数据，有 42809 条样本，17 个字段，包括了顾客编号，销售日期， 商品类型等 17 个字段。请根据 supermarket_sales.xlsx 提供的数据完成以下操作。

实施步骤：

任务一：根据"销售月份"、"销售金额"两列，绘制该超市各月销售金额占比饼图（matplotib.pyplot pie），并进行简单分析。

① 统计各月销售金额，统计方式参考 data.groupby（'销售月份'）.agg（{销售金额'：sum}；

② 设置饼图大小和百分比：plt.figure（figsize=（(5，5）；

③ 调用 pl.pie 绘制饼图，语法参考 pit，pie（数据,，labels=**，autopct=***）；

④ 设置标题（title）为'该超市各月销售金额占比饼图'；

⑤显示图表；

⑥ 简单分析各月销售金额情况；

任务二：根据"是否促销"、"销售月份"、"销售金额"三列，绘制促销商品月销售金额柱状图（matplotlib.pyplot.bar），并进行简单分析。

① 统计促销商品月销售金额，统计方式参考 data[data['是否促销]--'是]groupby（销售月份"）.agg（{'销售金额'sum}）；

② 调用 plt.bar 绘制柱状图，语法参考 pit.bar（数据范围，数据）；

③ 设置 x 轴（xticks）刻度，语法参考 plt.xticks（数据范围，数据行索

引，rotation=45）；

④ 设置 x 轴标签（xlabel）为"销售月份";

⑤ 设置 y 轴标签（ylabel）为"月销售金额";

⑥ 设置标题（tite）为"促销商品月销售金额柱状图";⑦ 显示图表;

⑧ 简单分析促销商品月销售金额情况。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

见附录 3：Python 数据可视化评分标准

（4）实施条件

见附录 9：数据可视化模块实施条件

## 2.Z1-2，招聘信息数据可视化与分析

（1）任务描述

zhaopin.xlsx 提供了某招聘网址 2019 年 10 月 23 日至 10 月 26 日发布的"数据分析"岗位招聘详情信息,用于挖掘并归纳出社会用人单位对数据分析师职位的招聘相关要求，以及招聘现状。zhaopin.xlsx 有 1737 条样本，14 个字段，包括了职位名，公司名，薪资 工作经验要求等 14 个字段。请根据 zhaopin.xlsx 提供的数据完成以下操作。

实施步骤：

任务一：根据"学历"一列，绘制一个展 示各学历的岗位需求量占比饼 图（matplotib.pyplot.pie），并进行简单分析。

① 统计各学历的岗位需求量，统计方式参考 data['学历]value_counts

② 设置画布 plt figure（figsize=（5）；

③ 调用 plt.pie 绘制饼图，语法参考 plt.pie（数据，labels=**，autopct=**）；

④ 设置标题（title）为'各学历的岗位需求量占比饼图';

⑤ 显示图表

⑥ 简单分析各学历的岗位需求量情况;

任务二：根据"工作经验要求"一列，绘制一个展示工作经验与岗位需求量关系的柱状图（ matplotlib.pyplot.bar ），并进行简单分析。

① 统计岗位对不同工作经验的需求量，统计方式参考 data[工作经验要求'].value_counts().sort_index(

② 调用 plt bar 绘制柱状图，语法参考 plt.bar（数据范围，数据）；

③ 设置 x 轴（xticks）刻度，语法参考 pltxticks（数据范围，数据行索引 rotation=45）；

④ 设置 x 轴（xlabel）为"工作经验";

⑤ 设置 y 轴（ylabel）标签为"需求量";

⑥ 循环添加数据标签：for ij in zip（(数据范围，数据）：

plt text(i, j, '%d'%, ha='center', va='bottom')

⑦ 添加标题为"经验的岗位需求量"

⑧ 显示图表

⑨ 简单分析岗位对不同工作经验的需求量情况;

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

见附录 3：Python 数据可视化评分标准

（4）实施条件

见附录 9：数据可视化模块实施条件

## 3.Z1-3，豆瓣网图书数据可视化与分析

（1）任务描述

book.xlsx 包含豆瓣网图书的书名，作者 出版社，出版时间，评分等数据。请根据 book.xlsx 数据完成以下操作。

实施步骤：

任务一：根据“作者‘一列，绘制作品数量最多的前 10 位作者排名柱状图。

①提取作品数量最多釣前十位作者排名数据.参考方式如下:

data.groupby(by-,作者).size。. sort_vaiues(ascending=False)[ 1:11 ]

②调用 ph.bar 绘制柱状图，语法参考 plt.bar(数据范匿，数据)；

③设置轴刻度(xticks),语法参考 pltxticks(数据范匿，数据行索 51,rotation=45) ®设置 x 轴标签（xlabel）作者

④ 没置 y 轴标签(viable):作品数量

⑤设置标题(title):作品数量最多的前十位作者排名柱状图

⑥显示柱状图

⑧简单分析作者排名情况。

任务二：根据"学历"一列，绘制一个展示各学历的岗位需求量占比饼图（matplotib.pyplot.pie），并进行简单分析。

① 统计各学历的岗位需求量，统计方式参考 data['学历]value_counts

② 设置画布 plt figure（figsize=（5）；

③ 调用 plt.pie 绘制饼图，语法参考 plt.pie（数据，labels=**，autopct=**）；

④ 设置标题（title）为'各学历的岗位需求量占比饼图';

⑤ 显示图表，简单分析各学历的岗位需求量情况;

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

见附录 3：Python 数据可视化评分标准

（4）实施条件

见附录 9：数据可视化模块实施条件

## 4.Z1-4，豆瓣影评数据可视化与分析

（1）任务描述

doubm.xlsx 是《哪吒》在豆瓣平台上的热门影评数据，包括了短评正文、评分、赞同数量等 7 个字段，主要用来分析《哪吒》该影片评论及观影群众相关信息。

实施步骤：

任务一：根据"评分"一列，绘制《哪吒》豆蓑评分分布饼图(matplotlib.pyplot.pie),并进 行简单分析。

① 统计各评分分布。

② 调用 pit* 绘制饼图。

③ 设置饼图标题(title)为“《哪吒》豆瓣评分分布饼图”。

④ 显示饼图，简单分析流浪地球评分分布。

任务二：根据“居住城市"一列，绘制评论数量最多的前 5 个城市排名柱状图 （matplotiib.pyplot.bar),并进行简单分析注：评论数量最多的前 5 个城市数据提取方式参考如下 data「居住城市'].value_counts()[:5]。

① 统计评论数量最多的前 5 个城市，统计方式参考 data「居住城市 Jvalue_coums()[:5]。

② 调用 plt.bar 绘制柱状图，语法参考 plt.bar(数据范围,数据)。

③ 设置 x 轴刻度(xticks),语法参考 plt.xticks(数据范匿，数据行索引,rotation=45)。

④ 设置 x 轴标签(xlabel)为“城市"。

⑤ 设置 y 轴标签(ylabel)为"评论数量。

⑥ 设置柱状图标题(title)为"评论数是最多的前 5 个城市排名柱状图。

⑦ 简单分析评论数最多的前 5 个城市情况。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

见附录 3：Python 数据可视化评分标准

（4）实施条件

见附录 9：数据可视化模块实施条件

## 5.Z1-5，票房数据可视化与分析

（1）任务描述

现有文件 emoxies.csv 包括了知名度，票房，名 称等 16 个字段，通过对该文件数据进行清洗，筛选，通过图形控件展示统计分析结果。

| 字段名称 | 字段说明 |
| --- | --- |
| id | |
| imdb.id | IMDB 号 |
| popularity | 知名度 |
| budget | 预算 |
| revenue | 票房 |
| originaLtitle | 名称 |
| cast | 主演 |
| director | 导演 |
| overview | 简介 |

| | |
|---|---|
| runtime | 时长 |
| genres | 类别 |
| release_date | 发行日期 |
| vote_count | 投票总数 |
| vote_average | 投票均值 |
| release_year | 发行年份 |
| profit | 净利润 |

实施步骤：

任务一：根据“popularity”、Priginal_title”两列，绘制知名度最高的前 10 剖电影排名饼图,并进行简单分析。统计知名度排在前 10 的电影数据提，取方式参考如下：

data[['popularit,','original_title']].sort__values('popularitxr'; ascending=False)[: 10]

① 调用 pit* 绘制饼图。

② 设置饼图标题(title)为“《流浪地球》豆瓣评分分布饼图”。

③ 显示饼图，简单分析流浪地球评分分布。

任务二:根据“类别"一列，绘制票房数量最多的前 5 个类别排名柱状图,并进行简单分析注：票房数量最多的前 5 个类别数据提取方式参考如下 data「类别'].value_counts()[:5]。

① 统计票房数量最多的前 5 个类别。

② 调用 plt.bar 绘制柱状图，语法参考 plt.bar(数据范围,数据)。

③ 设置 x 轴刻度(xticks),语法参考 plt.xticks(数据范匿，数据行索引,rotation=45)。

④ 设置 x 轴标签(xlabel)为“类别"。

⑤ 设置 y 轴标签(ylabel)为"票房数量。

⑥ 设置柱状图标题(title)为"票房数是最多的前 5 个类别排名柱状图。

⑦ 简单分析票房数最多的前 5 个类别情况。

作品提交：创建“所属学校_身份证_姓名_题号”命名的 word 文档，根据步骤截取对应操作命令与执行结果，所有代码和文档一起打包到“所属学校_身份证_姓名_题号.rar”文件中。

（2）考核时量

考核时间为 3 个小时。

（3）评分标准

见附录 3：Python 数据可视化评分标准

（4）实施条件

见附录 9：数据可视化模块实施条件

# 附录 1：算法设计与实现评分标准

| 类别 | 适用项 | 评价项 | 分值 | 评分标准 |
|---|---|---|---|---|
| 技能要求（90 分） | 任务一（30）<br>任务二（30）<br>任务三（30） | 开发环境使用正确性 | 5 分 | 按要求提交正确格式源文件，5 分 |
| | | 流程图设计合理性 | 10 分 | 流程图逻辑不正确扣 10 分；流程逻辑正确符号不当，每个符号扣 2 分扣完为止 |
| | | 程序设计合理性 | 5 分 | 程序中出现无用变量，非必要循环，分支结构扣 1 分一个，扣完为止 |
| | | 功能实现 | 10 分 | 按照任务要求实现相应功能 10 分 |
| 素养要求（10 分） | 整体 | 代码书写规范 | 3 分 | 代码缩进不规范扣 1 分、方法定义不规范扣 1 分、语句结构不规范扣 1 分 |
| | | 注释规范 | 2 分 | 无注释扣 2 分，注释不规范扣 1 分 |
| | | 命名规范 | 5 分 | 类名，变量名，方法名命名不规范每一个扣 1 分，扣完为止 |

# 附录 2：数据库设计评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | 创建数据库 | 5 分 | 正确创建数据库 5 分 |
| | 创建表 | 30 分 | 1. 创建表 1,表 2 设置对应字段类型正确 2 分一项，总：20 分<br>2. 设置对应主键自增 4 分，外键关系 6 分 |
| | 添加操作 | 10 分 | 每个表添加 5 条数据：1 条 1 分 总分：10 分 |
| | 更新操作 | 10 分 | 按要求完成 4.2 更新操作： 10 分 |
| | 查询操作 | 35 分 | 1、按要求完成 4.3 操作：10 分<br>2、按要求完成 4.4 操作：10 分<br>3、按要求完成 4.5 操作：15 分 |
| 素养要求 | 命名规范 | 4 分 | 1、数据库命名正确： 2 分<br>2、表命名规范正确： 2 分 |
| | SQL 编写规范 | 6 分 | 1、缩进合理:3 分<br>2、有对应注释：3 分 |

## 附录 3：Python 数据可视化评分标准

| 类别 | 评价项 | 分值 | 评分标准 |
| --- | --- | --- | --- |
| 技能要求 | Python 语法基础使用 | 25 分 | 正确导入 pandas 5 分<br>正确导入 matplotlib.py 的 5 分<br>正确读取文件 10 分<br>正确设置中午展示 5 分 |
| | 绘制饼图 | 30 分 | 饼图绘制正确得 10 分<br>饼图包含标题，占比比例显示，项目名称显示等设置 10 分<br>最高月份正确 5 分<br>最低月份正确 5 分 |
| | 绘制柱状图 | 35 分 | 柱状图正确 15 分<br>包含标题，轴标签，轴刻度 10 分<br>最高月份正确 10 分 |
| | 文档规范 | 10 分 | 正确提交文档，截图完整，结构清晰 5 分<br>代码规范，有注释 5 分 |

## 附录 4：程序设计模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
| --- | --- | --- | --- |
| 1 | 计算机 | 双核 CPU，内存 4G 或以上，win10 操作系统 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | JDK1.8 或以上，Python3.0 或以上 | Java 和 Python 开发环境 | |
| 4 | Intelij IDEA2019 或以上，Pycharm2019 或以上 | 软件开发 | 参考人员自选开发工具 |
| 5 | MSDN 或 JDK 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |

## 附录 5：数据库设计模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
| --- | --- | --- | --- |
| 1 | 计算机 | 双核 CPU，内存 4G 或以上，win10 操作系统 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | MySQL5.7 或以上<br>Navicat | 数据库管理系统 | |

## 附录 6：网络爬虫模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
| --- | --- | --- | --- |
| 1 | 计算机 | 双核 CPU，内存 8G 或以上，win10 操作系统 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | Python3.0 或以上，爬虫相关库 | 爬虫环境 | |
| 4 | Intelij IDEA2019 或以上，<br>Pycharm2019 或以上 | 软件开发 | 参考人员自选开发工具 |
| 5 | Python 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |
| 6 | 网络环境 | 指定对应网站开启互联网环境/或提供本地网站资源环境 | |

## 附录 7：Hadoop 平台与组件模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
| --- | --- | --- | --- |
| 1 | 计算机 | 四核 CPU，内存 8G 或以上，win10 操作系统<br>或者同等条件的云服务器 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | VMware Workstation Pro12 及以上<br>Xshell,tabby,electerm 等 SSH 连接工具 | 虚拟机<br>SSH 工具 | 虚拟机 Hadoop 环境完备。<br>参考人员自选工具 |
| 4 | JDK 帮助文档<br>Hadoop 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |
| 5 | 网络环境 | 指定对应网站开启互联网 | |

## 附录 8：数据分析模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
| --- | --- | --- | --- |
| 1 | 计算机 | 四核 CPU，内存 16G 以上，win10 操作系统<br>或者同等条件的云服务器 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | VMware Workstation Pro12 及以上（安装好对应 hadoop 及 hive 环境）<br>Xshell,tabby,electerm 等 SSH 连接工具 | 虚拟机<br>SSH 工具 | 多台虚拟内 hadoop,Flink,Spark 相关环境完备。<br>参考人员自选连接工具 |
| 4 | JDK 帮助文档<br>Spark,Flink 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |
| 5 | 网络环境 | 指定对应网站开启互联网 | |

# 附录 9：数据可视化模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
| --- | --- | --- | --- |
| 1 | 计算机 | 四核 CPU，内存 8G 或以上，win10 操作系统<br>或者同等条件的云服务器 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | Python3.0 或以上<br>（安装 pyecharts、matplotlib 等图形库） | 开发环境 | |
| 4 | Intelij IDEA2019 或以上，<br>Pycharm2019 或以上 | 软件开发 | 参考人员自选开发工具 |
| 5 | MSDN 或 JDK 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |
| 6 | 数据资源 | 本地数据集 | |