



永州职业技术学院
Yongzhou Vocational Technical College

永州职业技术学院

大数据应用学生专业技能考核题库

专业代码:	510201
专业名称:	计算机应用技术 (大数据应用)
二级学院:	信息工程学院

永州职业技术学院
2021年6月

本专业技能考核以省教育厅《关于开展 2022 年高职高专院校专业人才培养方案、专业技能考核标准与题库、新设专业办学水平合格性评价和学生专业技能抽查工作的通知》为编制依据，通过设置大数据开发基础模块、数据采集与存储模块、数据清洗与分析应用模块、数据可视化模块等 4 个技能考核模块，测试学生的编程能力、数据采集能力、大数据存储能力、数据清洗能力、数据分析能力、数据可视化能力、项目管理能力以及从事大数据技术工作的团队协作、成本控制、质量效益、安全规范等职业素养。引导学校加强专业教学基本条件建设，深化课程教学改革，强化实践教学环节，增强学生创新创业能力，促进学生个性化发展，提高专业教学质量和专业办学水平，培养适应信息时代发展需要的大数据技术高素质技术技能人才。

题目设计以企、事业单位应用项目为背景，完成项目开发平台的配置与使用、项目模型的设计与建立、程序代码的编写与运行等工作内容。根据《计算机应用技术（大数据应用）人才培养方案》中关于专业面向的职业岗位与职业能力要求，设计本题库。本题库分为四大模块，包括大数据开发基础模块、数据采集与存储模块、数据清洗与分析模块、数据可视化模块。大数据开发基础模块，包括编程基础、数据库基础、平台部署与基础应用等 3 个项目内容；数据采集与存储模块，包括 flume 数据采集、kafka 数据采集、HBase 数据存储等 3 个项目内容；数据清洗与分析模块，包括 pandas 数据清洗与分析、数据 ETL 处理、Hive 数据分析、spark 数据分析等 4 个项目；数据可视化，包括 matplotlib 数据可视化、pyecharts 数据可视化等 2 个项目内容。题库内容基本涵盖了大数据运维工程师、大数据采集与处理工程师、大数据分析与可视化开发工程师、数据库工程师等岗位从事项目设计与开发工作所需的基本技能与能力要求。

大数据开发基础模块，包括用 Python、Java 语言（任一）编程解决实际问题、MySQL 数据库开发、hadoop 完全分布式平台部署、使用 HDFS 对 hadoop 平台进行基础开发。本模块涵盖了大数据运维工程师所需的编程基础、平台部署和开发的基础与核心技能。

数据采集与存储模块，包括以大数据技术 flume、kafka 实现数据采集，以 HBase 数据库实现非结构化数据存储。其中 flume 技术采集端口数据、文件数据、

变化数据，kafka采集实时的消息。本模块涵盖了大数据采集与处理工程师所需的数据采集与存储的核心技能。

数据清洗与分析模块，包括pandas数据清洗与分析、使用kettle实现数据ETL操作、hive数据分析、spark数据分析，实现批量与实时数据的清洗处理与分析。本模块基本涵盖了大数据采集与处理工程师所需的数据清洗、数据ETL核心技能，涵盖了大数据分析可视化开发工程师所需的批量数据、实时数据分析的核心技能。

数据可视化模块，包括以Python、pycharm为工具，实现matplotlib、pyecharts数据可视化操作，将实际应用中的各种不同类型的数据形成图表进行展示。本模块基本涵盖了大数据可视化工程师岗位进行数据展示核心技能。

目 录

模块一 大数据开发基础	1
项目 1: 编程基础	1
1. 试题编号: 1-1	1
2. 试题编号: 1-2	4
3. 试题编号: 1-3	7
4. 试题编号: 1-4	10
5. 试题编号: 1-5	13
项目 2: 数据库基础	16
6. 试题编号: 1-6	16
7. 试题编号: 1-7	21
项目 3: 大数据平台部署与基础开发	错误! 未定义书签。
8. 试题编号: 1-8	错误! 未定义书签。
9. 试题编号: 1-9	错误! 未定义书签。
10. 试题编号: 1-10	错误! 未定义书签。
模块二 数据采集与存储	1
项目 1: 基于 Flume 的数据采集	38
11. 试题编号: 2-1	38
12. 试题编号: 2-2	43
13. 试题编号: 2-3	48
项目 2: 基于 kafka 的消息队列数据采集	错误! 未定义书签。
14. 试题编号: 2-4	错误! 未定义书签。
15. 试题编号: 2-5	错误! 未定义书签。
模块三 数据清洗与分析	38
项目 1: 基于 pandas 的数据清洗	错误! 未定义书签。
19. 试题编号: 3-1	错误! 未定义书签。
20. 试题编号: 3-2	错误! 未定义书签。
21. 试题编号: 3-3	错误! 未定义书签。
22. 试题编号: 3-4	错误! 未定义书签。
项目 2: 基于 kettle 的数据清洗	错误! 未定义书签。
23. 试题编号: 3-5	93
24. 试题编号: 3-6	97
25. 试题编号: 3-7	102
模块四 数据可视化	77
项目 1: matplotlib 数据可视化	151
41. 试题编号: 4-1	151
42. 试题编号: 4-2	155
43. 试题编号: 4-3	160
44. 试题编号: 4-4	164
45. 试题编号: 4-5	168
项目 2: pyecharts 数据可视化	172
46. 试题编号: 4-6	172
47. 试题编号: 4-7	176

48. 试题编号: 4-8.....	180
49. 试题编号: 4-9.....	184
50. 试题编号: 4-10.....	188

模块一 大数据开发基础

项目 1：编程基础

请使用 Python、Java 中任一编程语言，进行编程设计。

1. 试题编号：1-1

任务一 打印 9*9 乘法口诀表（40 分）

（1）任务描述

编程实现控制台打印九九乘法口诀。

（2）任务要求

1. 根据题目描述，编写程序。
2. 正确调试程序。

任务二 编写一个函数，实现接收一个列表，返回这个列表的最大值、平均值、最小值的功能（40 分）

（1）任务描述

编写代码定义一个函数，函数参数为列表，函数功能为求列表的最大值，平均值，最小值，函数返回值为列表最大值、平均值、最小值。编写主函数，定义一个列表，应用函数。

（2）任务要求

1. 根据任务描述与分析，编写程序。
2. 正确调试程序。

提交要求：

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 1-1-1 模块项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019、ideal2019 或更高版本	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

Python 程序设计模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 1-1-2 考核评价标准

评价内容		评分标准		备注
工作任务一	9*9 口诀	输出的 9*9 口诀是否符合要求	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶劣影响的本项目记 0
	变量使用	变量声明是否符合要求	5 分	
	循环结构书写	循环结构书写是否符合要求	5 分	
工作任务二	定义函数分析列表数据	最大值、最小值、平均值	30 分	
	变量使用	变量声明是否符合要求	5 分	

	主函数	主函数语句是否符合要求	5分	分。
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
合计		100分		

2. 试题编号：1-2

任务一 输入两个正整数求最大公约数和最小公倍数（40分）

（1）任务描述

编写程序，实现学生从键盘输入两个数，得出这两个数的最大公约数和最小公倍数的功能。

（2）任务要求

1. 根据任务描述与分析，编写程序。
2. 正确调试程序。

任务二 判断回文串（40分）

（1）任务描述

回文串是指这个字符串无论从左读还是从右读，都是相同的。请编写一个程序，接收用户输入的一个字符串，然后判断它是否是回文串。

（2）任务要求

1. 根据任务描述与分析，编写程序。
2. 正确调试程序。

提交要求

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 1-2-1 项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本	用于程序设计，每人一台。

	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019、ideal2019 或更高版本	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

Python 程序设计模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 1-2-2 考核评价标准

评价内容		配分	评分标准		备注
工作任务一	最大公约数、最小公倍数计算	30 分	输出正确最大公约数、最小公倍数	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶劣影响的本项目记 0 分。
	辗转相除法	5 分	辗转相除法	5 分	
	判断、循环语句	5 分	判断是否符合要求	5 分	
工作任务二	判断回文串	30 分	正确判断	30 分	
	变量使用	5 分	变量声明是否符合要求	5 分	

	判断语句	5 分	判断是否符合要求	5 分	
职业素养	专业素养	10 分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	10 分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分			

3. 试题编号：1-3

任务一 求年龄（40分）

（1）任务描述

有五个人坐在一块，问第五个人多少岁？条件如下：

第五个人说他比第四个人大2岁。第四个人说他比第三个人大两岁。以此类推，问道最后一个人（第一个人），他说自己十岁。

（2）任务要求

1. 根据题目描述，编程程序。
2. 正确调试程序。

任务二 空气质量等级（40分）

（1）任务描述

空气质量问题一直是社会所关注的，一种简化的判别空气质量的方式如下：PM2.5的数值为0~35(包括0但不包括35)为优，35~75(包括35和75)为良，75以上为污染。请编写程序实现如下功能：输入PM2.5的值，输出空气质量情况。

（2）任务要求

1. 根据题目描述，编程程序。
2. 正确调试程序。

提交要求：

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 1-3-1 实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机	用于程序设

	安装 Windows 7 或更高版本		计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019、ideal2019 或更高版本	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

Python 程序设计模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 1-3-2 考核评价标准

评价内容		分值	评分标准		备注
工作任务一	求年龄	30 分	正确分析年龄规律，求出年龄	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶劣影响的本项目记 0 分。
	变量定义	5 分	合理定义变量	5 分	
	循环语句	5 分	判断是否符合要求	5 分	
工作任务二	空气质量等级	30 分	正确判断数值分布的区间，结果输出正确	30 分	

	判断语句	5分	正确使用判断语句	5分	
	变量定义	5分	合理定义变量	5分	
职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

4. 试题编号：1-4

任务一 求 1~100 的累加值（40 分）

（1）任务描述

打印 1+2+3+4+5+6+···+100 的累加值

（2）任务要求

1. 根据题目描述，编程程序。
2. 正确调试程序。

任务二 实现数学方程式（40 分）

（1）任务描述

编写程序，实现下面的方程：

$$y = \begin{cases} 0, & x < 5 \\ 5 \times x - 25, & 5 \leq x < 10 \\ (x - 5)^2, & 10 \leq x \end{cases}$$

输入数据，进行验证

（2）任务要求

1. 根据题目描述，编程程序。
2. 正确调试程序。

提交要求：

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 1-4-1 模块项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本	用于程序设计，每人一台。

	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019、ideal2019 或更高版本	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

Python 程序设计模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 1-4-2 考核评价标准

评价内容		分值	评分标准		备注
工作任务一	1~100 的累加	30 分	结果输出正确	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶劣影响的本项目记 0 分。
	变量定义	5 分	合理定义变量	5 分	
	循环使用	5 分	循环正确使用	5 分	
工作任务二	实现数学方程式	30 分	结果输出正确	30 分	
	判断使用	5 分	判断语句使用正确	5 分	

	运算符的使用	5分	运算符正确使用	5分	
职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

5. 试题编号：1-5

任务一 整除问题（40分）

（1）任务描述

编写程序，功能如下：判断输入的一个整数能否同时被 2 和 3 整除，若能，则输出 “Yes”；否则输出 “No”。

（2）任务要求

1. 根据题目描述，编程程序。
2. 正确调试程序。

任务二 本金与利息问题（40分）

（1）任务描述

本金 10000 元存入银行，年利率是 2%。每过 1 年，将本金和利息相加作为新的本金。要求计算 5 年后的本金。

（2）任务要求

1. 根据题目描述，编程程序。
2. 正确调试程序。

提交要求：

1) 在 “e:\技能抽查提交资料\” 文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以 “姓名_题号.py” 命名，最终将考生文件夹进行压缩后提交。

表 1-5-1 模块项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本	用于程序设计，每人一台。

	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019、ideal2019 或更高版本	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

Python 程序设计模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 1-5-2 考核评价标准

评价内容		分值	评分标准		备注
工作任务一	整除	30 分	结果输出正确	30 分	
	变量定义	5 分	变量定义合理	5 分	
	判断结构	5 分	判断语句正确使用	5 分	
工作任务二	本金与利息计算	30 分	结果输出正确	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本
	循环语句	5 分	循环语句使用正确	5 分	

	每年的本金计算	5分	每年的本金计算正确	5分	项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

项目 2：数据库基础

6. 试题编号：1-6

(1) 任务描述

《网上商店》的 E-R 图如图 1.6.1 所示；逻辑数据模型如图 1.6.2 所示；数据表字段名定义见表 1-6-1，表数据分别建表 1-6-2、表 1-6-3、表 1-6-4。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：

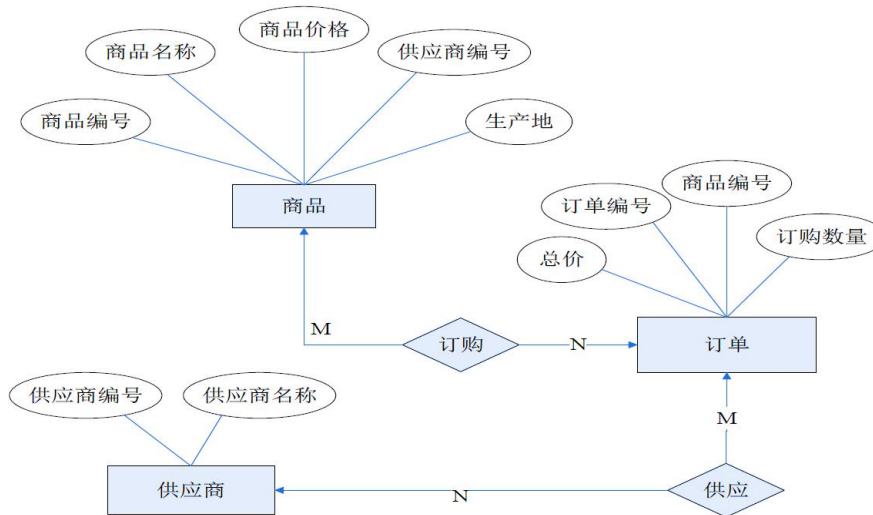


图 1.6.1 E-R 图

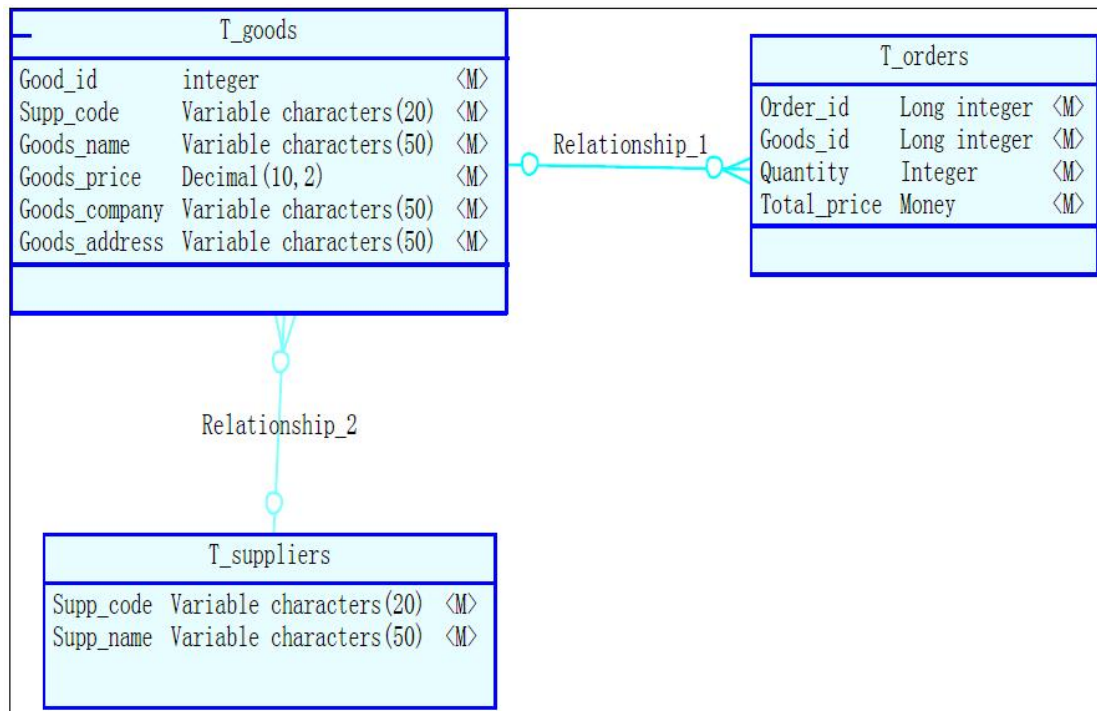


图 1.6.2 逻辑数据模型

表 1-6-1 字段名定义表

字段名	字段说明	字段名	字段说明
Goods_id (标识列)	商品编号	Quantity	订购数量
Goods_name	商品名称	Total_price	总价
Goods_price	商品价格	Supp_code	供应商编号
Gupp_code	供应商编号	Supp_name	供应商名称
Goods_adress	生产地		
Order_id	订单号码		
Goods_id	商品编号		

表 1-6-2 商品信息表 (T_goods 样本数据)

Goods_id	Goods_name	Goods_price	Supp_code
1000	盛唐笔记本	5600	430102
1001	博士笔记本	6700	540199
1002	惠普笔记本	7800	440708

表 1-6-3 订单信息表 (T_orders 样本数据)

Order_id	Goods_id	Quantity	Total_price
11070232	1000	3	16800
11060343	1002	1	7800
11050322	1001	2	13400

表 1-6-4 供应商表 (T_supplies 样本数据)

Supp_code	Supp_name
430102	盛唐科技
540199	博士科技
440708	惠普科技

任务一 DDL 操作：创建数据库和数据表（共 15 分）

- (1) 创建数据库 Stores（可设置 charset 为 utf8）。（3 分）

- (2) 创建数据表
 - a) 根据表 1-6-2，在 Stores 数据库中，创建数据表 T_goods。（4 分）
 - b) 根据表 1-6-3，在 Stores 数据库中，创建数据表 T_orders。（4 分）
 - c) 根据表 1-6-4，在 Stores 数据库中，创建数据表 T_supplies。（4 分）

任务二 DML 操作（共 35 分）

- (1) 根据表 1-6-2，向数据表 T_goods 中插入数据。（9 分）

- (2) 根据表 1-6-3，向数据表 T_orders 中插入数据。（9 分）

- (3) 根据表 1-6-4，向数据表 T_supplies 中插入数据。（9 分）

- (4) 将商品名为“惠普笔记本”的价格下调 10%；（8 分）

任务三 DQL 操作（每题 5 分，共 30 分）

- (1) 查询出商品编号为“1002”的订购数量。

- (2) 查询出商品编号为“1000”的商品名称和商品价格。

- (3) 查询出所有商品的总订购数量。

- (4) 从商品信息表中，统计商品的种类。

- (5) 查询出商品名称为“惠普笔记本”的商品的订购数量、总价；

(6) 查询所有名称包含“科技”的供应商编号、供应商名称。

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 1-6 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX， 并给出运行结果截图

(3) 实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	MySQL、Navicat	用以 SQL 语句开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项

目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	DDL 操作	数据库创建、数据表创建	15 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	DML 操作	数据插入正确，少插入一条数据扣 3 分	35 分	
	DQL 操作	数据查询正确	30 分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分		

7. 试题编号：1-7

(1) 任务描述

《网上商店》的 E-R 图如图 1.7.1 所示；逻辑数据模型如图 1.7.2 所示；数据表字段名定义见表 1-7-1，表数据分别建表 1-7-2、表 1-7-3、表 1-7-4。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：

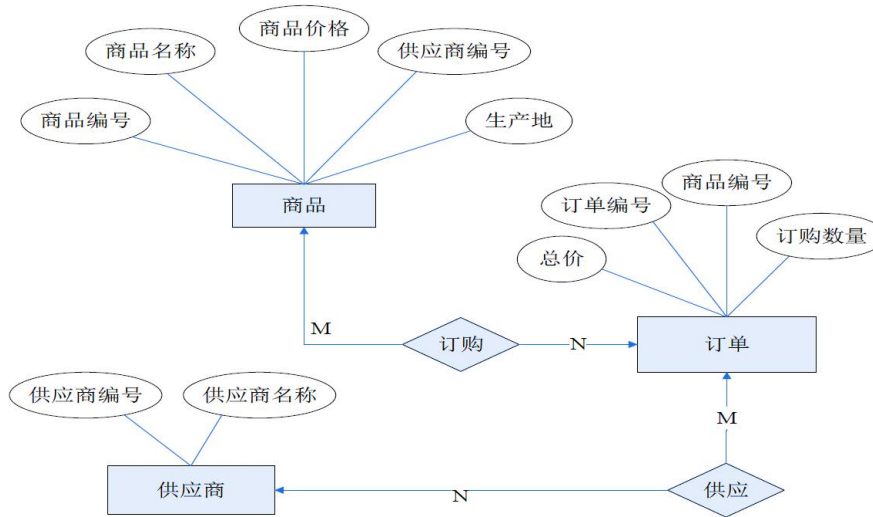


图 1.7.1 E-R 图

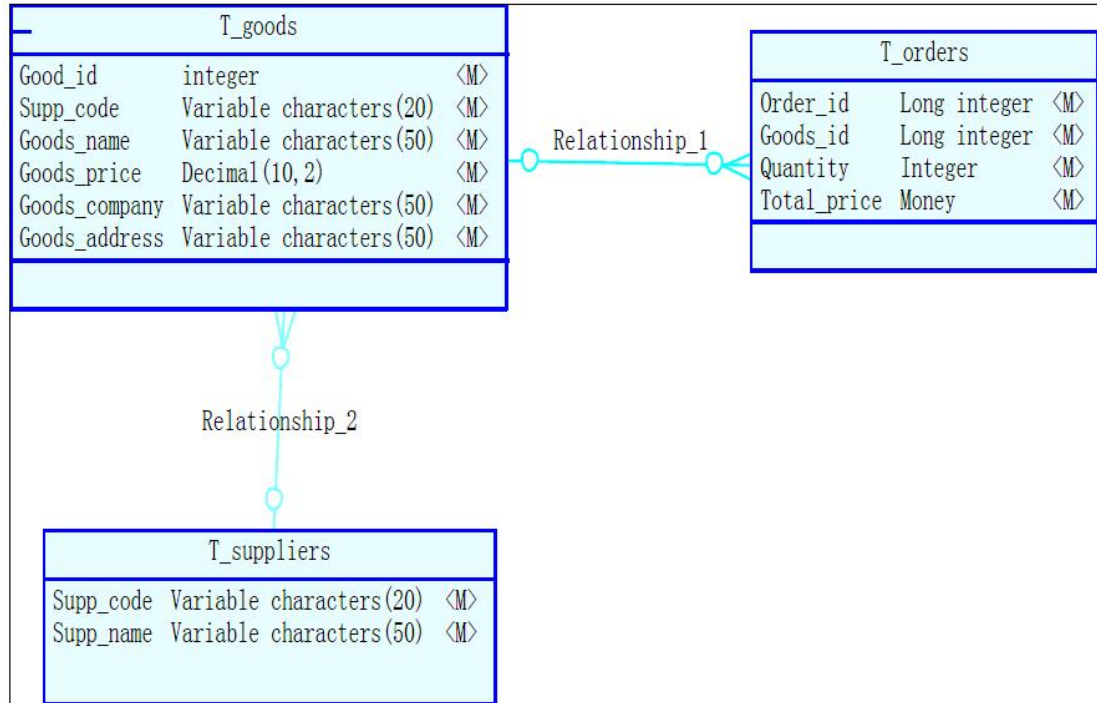


图 1.7.2 逻辑数据模型

表 1-7-1 字段名定义表

字段名	字段说明	字段名	字段说明
-----	------	-----	------

Goods_id (标识列)	商品编号	Quantity	订购数量
Goods_name	商品名称	Total_price	总价
Goods_price	商品价格	Supp_code	供应商编号
Gupp_code	供应商编号	Supp_name	供应商名称
Goods_adress	生产地		
Order_id	订单号码		
Goods_id	商品编号		

表 1-7-2 商品信息表 (T_goods 样本数据)

Goods_id	Goods_name	Goods_price	Supp_code
1000	盛唐笔记本	5600	430102
1001	博士笔记本	6700	540199
1002	惠普笔记本	7800	440708

表 1-7-3 订单信息表 (T_orders 样本数据)

Order_id	Goods_id	Quantity	Total_price
11070232	1000	3	16800
11060343	1002	1	7800
11050322	1001	2	13400

表 1-7-4 供应商表 (T_supplies 样本数据)

Supp_code	Supp_name
430102	盛唐科技
540199	博士科技
440708	惠普科技

任务一 DDL 操作：创建数据库和数据表（共 15 分）

- (1) 创建数据库 Stores（可设置 charset 为 utf8）。（3 分）

(2) 创建数据表

- a) 根据表 1-7-2, 在 Stores 数据库中, 创建数据表 T_goods。 (4 分)
- b) 根据表 1-7-3, 在 Stores 数据库中, 创建数据表 T_orders。 (4 分)
- c) 根据表 1-7-4, 在 Stores 数据库中, 创建数据表 T_supplies。 (4 分)

任务二 DML 操作 (共 35 分)

- (1) 根据表 1-7-2, 向数据表 T_goods 中插入数据。 (9 分)
- (2) 根据表 1-7-3, 向数据表 T_orders 中插入数据。 (9 分)
- (3) 根据表 1-7-4, 向数据表 T_supplies 中插入数据。 (9 分)
- (4) 将商品名为“盛唐笔记本”的价格上调 10%; (8 分)

任务三 DQL 操作 (每题 5 分, 共 30 分)

- (1) 查询出商品价格大于 6000 的商品编号、商品名称。
- (2) 查询出订购数量最多的商品编号。
- (3) 查询出所有商品的总订购数量。
- (4) 查询出订购数量最少的商品编号、总价和商品名称。
- (5) 查询出商品价格大于 7000 的商品名称、供应商名称。
- (6) 查询商品名称包含“惠普”的商品编号、商品价格、供应商编号。

将该答案文件保存到考生文件夹中。

提交要求:

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 1-7 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图: XXXXXX, 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	MySQL、Navicat	用以 SQL 语句开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成

情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	DDL 操作	数据库创建、数据表创建	15 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	DML 操作	数据插入正确，少插入一条数据扣 1 分	35 分	
	DQL 操作	数据查询正确	30 分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分		

项目 3：大数据平台部署与基础开发

8. 试题编号：1-8

(1) 任务描述

你作为公司某一项目开发人员，需要对项目代码进行部分功能测试，因此可以搭建伪分布式 hadoop 环境，节省资源并实现功能测试。本项目主要完成 Hadoop 伪分布式部署，本环节需要使用 centos 系统、root 用户完成相关配置。

任务一 搭建 JDK 环境（每题 2 分，共 8 分）

- (1) 解压路径/opt/software 下的 JDK 压缩包，到路径/opt/module
- (2) 解压成功后，更改名称为 jdk18
- (3) 配置 JDK 系统环境(/etc/profile)变量，并激活
- (4) 查看 JDK 版本信息，验证 JDK 是否成功安装

任务二 SSH 免密登录设置（每题 2 分，共 6 分）

- (1) 生成密钥对
- (2) 追加公钥
- (3) 免密登录验证

任务三 Hadoop 伪分布式搭建（共 16 分）

- (1) 解压路径/opt/software 下的 Hadoop 压缩包，到路径/opt/module
(1 分)
- (2) 解压成功后，更改名称为 hadoop (1 分)

- (3) 配置 hadoop 系统环境(/etc/profile)变量, 并激活 (1分)
- (4) 查看 Hadoop 版本信息, 验证 hadoop 是否成功安装 (1分)
- (5) 设置 hadoop 中的配置文件 hadoop-env.sh, 加入 Java 路径 (2分)
- (6) 设置 hadoop 中的配置文件 core-site.xml (2分)
- (7) 设置 hadoop 中的配置文件 hdfs-site.xml (4分)
- (8) 将 mapred-site.xml.tmp 文件复制为 mapred-site.xml, 设置 hadoop 中的配置文件 mapred-site.xml (2分)
- (9) 设置 hadoop 中的配置文件 yarn-site.xml (2分)

任务四 启动 Hadoop (每题 2 分, 10 分)

- (1) 格式化 HDFS
- (2) 启动 HDFS 进程
- (3) 启动 YARN 进程
- (4) 查看服务器进程, 验证 Hadoop 是否启动成功
- (5) 使用浏览器访问 HDFS, 验证 HDFS 是否启动成功

任务五 基础开发操作 (每题 5 分, 40 分)

- (1) 在 HDFS 上创建/test/hadoop 路径, 并进行验证

- (2) 上传服务器/opt/module/hadoop 路径下的 README.txt 文件，到 HDFS 的 /test/hadoop/ 路径下，并进行验证
- (3) 修改 HDFS 上的 README.txt 文件所属权限为 777
- (4) 统计 HDFS 上的 README.txt 文件大小信息
- (5) 查看 HDFS 上的 README.txt 文件末尾 1KB 的内容
- (6) 查看 HDFS 上的 README.txt 文件的副本数量
- (7) 设置 HDFS 上的 README.txt 文件的副本数量为 5
- (8) 将 HDFS 的 /test/hadoop/README.txt 文件移动到 HDFS 路径 /test 下

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 1-8 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Centos7 或更高版本	用于程序设计，每人一台。
	FTP 服务器 1 台	用于保存测试人员考试结果

工具	开发工具	XShell、SecureCRT、ideal	用以连接服务器、开发代码
测评专家	现场测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。		

(4) 考核时量

考核时间为90分钟

(5) 评分标准

数据采集模块的考核实行100分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的20%，工作任务完成质量占该项目总分的80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	搭建JDK环境	JDK的环境安装与配置、验证	8分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	SSH免密登录设置	免密登录配置	6分	
	Hadoop伪分布式搭建	Hadoop环境变量、配置文件配置、验证	16分	
	启动Hadoop	Hadoop各组件启动成功	10分	
	基础开发操作	基础开发操作正确	40分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分		

9. 试题编号：1-9

(1) 任务描述

你作为公司某一项目开发人员，需要对项目代码进行部分功能测试，因此可以搭建伪分布式 hadoop 环境，节省资源并实现功能测试。本项目主要完成 Hadoop 伪分布式部署，本环节需要使用 centos 系统、root 用户完成相关配置。

任务一 搭建 JDK 环境（每题 2 分，共 8 分）

- (1) 解压路径/opt/software 下的 JDK 压缩包，到路径/opt/module
- (2) 解压成功后，更改名称为 jdk18
- (3) 配置 JDK 系统环境变量，并激活
- (4) 查看 JDK 版本信息，验证 JDK 是否成功安装

任务二 SSH 免密登录设置（每题 2 分，共 6 分）

- (1) 生成密钥对
- (2) 追加公钥
- (3) 免密登录验证

任务三 Hadoop 伪分布式搭建（共 16 分）

- (1) 解压路径/opt/software 下的 Hadoop 压缩包，到路径/opt/module
(1 分)
- (2) 解压成功后，更改名称为 hadoop (1 分)
- (3) 配置 hadoop 系统环境变量，并激活 (1 分)
- (4) 查看 Hadoop 版本信息，验证 hadoop 是否成功安装 (1 分)

- (5) 设置 hadoop 中的配置文件 `hadoop-env.sh`, 加入 Java 路径 (2 分)
- (6) 设置 hadoop 中的配置文件 `core-site.xml` (2 分)
- (7) 设置 hadoop 中的配置文件 `hdfs-site.xml` (4 分)
- (8) 将 `mapred-site.xml.template` 文件复制为 `mapred-site.xml`, 设置 hadoop 中的配置文件 `mapred-site.xml` (2 分)
- (9) 设置 hadoop 中的配置文件 `yarn-site.xml` (2 分)

任务四 启动 Hadoop (每题 2 分, 10 分)

- (1) 格式化 HDFS
- (2) 启动 HDFS 进程
- (3) 启动 YARN 进程
- (4) 查看服务器进程, 验证 Hadoop 是否启动成功
- (5) 使用浏览器访问 HDFS, 验证 HDFS 是否启动成功

任务五 基础开发操作 (每题 5 分, 40 分)

- (1) 在 HDFS 上创建 `/test/hadoop` 路径, 并进行验证
- (2) 上传服务器 `/opt/module/hadoop` 路径下的 `README.txt` 文件, 到 HDFS 的 `/test/hadoop/` 路径下, 并进行验证

- (3) 将 HDFS 的 /test/hadoop/README.txt 文件复制到 HDFS 路径 /test 下
- (4) 下载 HDFS 的 /test/路径下的 README.txt 文件，到服务器路径 /opt
- (5) 删除 HDFS 上的 /test/README.txt 文件
- (6) 删除 HDFS 目录 /test/hadoop/
- (7) 运行 MapReduce 的案例 jar 包，计算圆周率 π ，设置 10 个 MAP 任务，每个 MAP 任务计算 10 次
- (8) 列出 YARN 的所有节点

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 1-9 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Centos7 或更高版本	用于程序设计，每人一台。
	FTP 服务器 1 台	用于保存测试人员考试结果

工具	开发工具	XShell、SecureCRT、ideal	用以连接服务器、开发代码
测评专家	现场测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。		

(4) 考核时量

考核时间为90分钟

(5) 评分标准

数据采集模块的考核实行100分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的20%，工作任务完成质量占该项目总分的80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	搭建JDK环境	JDK的环境安装与配置、验证	8分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	SSH免密登录设置	免密登录配置	6分	
	Hadoop伪分布式搭建	Hadoop环境变量、配置文件配置、验证	16分	
	启动Hadoop	Hadoop各组件启动成功	10分	
	基础开发操作	基础开发操作正确	40分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分		

10. 试题编号：1-10

(1) 任务描述

你作为公司某一项目开发人员，需要对项目代码进行部分功能测试，因此可以搭建伪分布式 hadoop 环境，节省资源并实现功能测试。本项目主要完成 Hadoop 伪分布式部署，本环节需要使用 centos 系统、root 用户完成相关配置。

任务一 搭建 JDK 环境（每题 2 分，共 8 分）

- (1) 解压路径/opt/software 下的 JDK 压缩包，到路径/opt/module
- (2) 解压成功后，更改名称为 jdk18
- (3) 配置 JDK 系统环境变量，并激活
- (4) 查看 JDK 版本信息，验证 JDK 是否成功安装

任务二 SSH 免密登录设置（每题 2 分，共 6 分）

- (1) 生成密钥对
- (2) 追加公钥
- (3) 免密登录验证

任务三 Hadoop 伪分布式搭建（共 16 分）

- (1) 解压路径/opt/software 下的 Hadoop 压缩包，到路径/opt/module
(1 分)
- (2) 解压成功后，更改名称为 hadoop (1 分)
- (3) 配置 hadoop 系统环境变量，并激活 (1 分)
- (4) 查看 Hadoop 版本信息，验证 hadoop 是否成功安装 (1 分)

- (5) 设置 hadoop 中的配置文件 `hadoop-env.sh`, 加入 Java 路径 (2 分)
- (6) 设置 hadoop 中的配置文件 `core-site.xml` (2 分)
- (7) 设置 hadoop 中的配置文件 `hdfs-site.xml` (4 分)
- (8) 将 `mapred-site.xml.tmp` 文件复制为 `mapred-site.xml`, 设置 hadoop 中的配置文件 `mapred-site.xml` (2 分)
- (9) 设置 hadoop 中的配置文件 `yarn-site.xml` (2 分)

任务四 启动 Hadoop (每题 2 分, 10 分)

- (1) 格式化 HDFS
- (2) 启动 HDFS 进程
- (3) 启动 YARN 进程
- (4) 查看服务器进程, 验证 Hadoop 是否启动成功
- (5) 使用浏览器访问 HDFS, 验证 HDFS 是否启动成功

任务五 基础开发操作 (每题 5 分, 40 分)

- (1) 在 HDFS 上创建 `/test/hadoop` 路径, 并进行验证
- (2) 上传服务器 `/opt/module/hadoop` 路径下的 `README.txt` 文件, 到 HDFS 的 `/test/hadoop/` 路径下, 并进行验证
- (3) 将 HDFS 的 `/test/hadoop/README.txt` 文件复制到 HDFS 路径 `/test` 下

- (4) 设置 HDFS 上的 README.txt 文件的副本数量为 5
- (5) 列出 YARN 的所有节点
- (6) 下载 HDFS 的/test/路径下的 README.txt 文件，到服务器路径/opt
- (7) 删除 HDFS 上的/test/README.txt 文件
- (8) 删除 HDFS 目录/test/hadoop/

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 1-10 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT、ideal	用以连接服务器、开发代码
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件

	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	
--	--	--

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	搭建 JDK 环境	JDK 的环境安装与配置、验证	8 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶劣影响的本项目记 0 分。
	SSH 免密登录设置	免密登录配置	6 分	
	Hadoop 伪分布式搭建	Hadoop 环境变量、配置文件配置、验证	16 分	
	启动 Hadoop	Hadoop 各组件启动成功	10 分	
	基础开发操作	基础开发操作正确	40 分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分		

模块二 数据采集与存储

项目 1：基于 Flume 的数据采集

11. 试题编号：2-1

(1) 任务描述

国内某 CDN 厂商，构建了全国性的服务网络，其为各大互联网厂商提供网络加速、内容分发等服务。其运营研发部构建了自动化运维系统，通过命令集或者脚本集来完成后台的运维日常工作。现在运营研发部为了监控这些脚本的运行效率以及对其进行优化，需要收集这些脚本或者命令的运行结果数据。现委托某工程师进行技术调研，其初步选定使用 Flume 框架来完成数据的采集，其调研的重点是产品能够收集大量脚本的运行结果数据到 Flume 中。

本项目主要完成结合任务描述完成 Flume 的组件选择、画出相应的 agent 拓扑图、编写其对应的配置文件、启动 agent 处理数据并使用命令验证数据是否正确处理等功能。

任务一 选择合适的 Flume 组件（总共 10 分）

- (1) 根据项目描述，选择能够支持处理指定文件的 Flume source 组件(4分)；
- (2) 根据项目描述，选择能够进行快速测试的 Flume channel 组件（3分）；
- (3) 根据项目描述，选择能够实时展示测试数据的 Flume sink 组件（3分）；

任务二 画出 agent 的拓扑图（总共 10 分）

根据任务一选取的 Flume source 和 channel、sink 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 a1。（10分）

任务三 编写 agent 的配置文件（总共 45 分）

前期准备：进入本地 flume 目录，新建 logs 文件夹和 job 文件夹。在 logs 文件夹下新建文件 1.log，键入 test 后保存。

- (1) 根据任务描述和拓扑图，在 job 文件夹下，创建 agent 的配置文件 a1_2_1.conf, 放到此文件夹。确定所使用的 Flume source 组件的别名、Flume c

hannel 组件的别名、Flume sink 组件的别名。(5 分)

- (2) 编写前面创建的配置文件，定义好整个 agent 所使用的组件。(5 分)
- (3) 编写 exec source 组件配置项：
 - a) 配置 source 组件的类型标识配置项(type); (5 分)
 - b) 配置 source 组件监听的 Unix 命令或者脚本(command)，监听 logs 文件夹中的 1.log; (4 分)
- (4) 编写 Memory channel 组件的配置项：
 - a) 配置 channel 组件的类型标识配置项(type); (5 分)
 - b) 配置 channel 组件容量大小配置项(capacity)为 1000; (4 分)
 - c) 配置 channel 组件事务容量大小配置项(transactionCapacity)为 100; (4 分)
- (5) 编写 Logger sink 组件的配置项：
 - a) 配置 sink 组件的类型标识配置项(type); (5 分)
- (6) 将创建好的 flume 组件组装为完整的 agent：
 - a) 配置 source 组件需要连接的 channel 的配置项(channels); (4 分)
 - b) 配置 sink 组件需要连接的 channel 的配置项(channel); (4 分)

任务四 使用 Flume 命令启动 agent 处理数据 (总共 5 分)

- (1) 进入 flume 目录，使用 Flume 的 bin 目录下 flume-ng 脚本，通过指定 flume 的配置文件所在目录(-c)、agent 的名称(-n)、agent 配置文件所在

的目录(-f)来启动编写的 Flumeagent，并且通过-Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台。（5分）

任务五 验证数据是否正确处理（总共 10 分）

- (1) 追加“hello bigdata”到 1.log；（5分）
- (2) 进行验证，logger sink 会将读取到指定文件的数据打印到控制台，将读取到的消息以及成功消费数据的截图存入到答案文件中。（5分）

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-1 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT、Hadoop 分布式系统	用以开发项目
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析		

	师、数据库设计师资格证书（2人/场）。	
--	---------------------	--

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注	
工作任务	选择正确 Flume Source、channel、sink	选择的 Flumesource、channel、sink 是否符合要求	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。	
	画出正确的拓扑图	根据选择的组件是否画出正确的 agent 图	10 分		
	编写 flume 配置文件	正确使用命令创建 agent 的配置文件 确定使用的 Flume 组件的别名 正确定义 agent 包含的组件			10 分
		正确配置 source			9 分
		正确配置 channel			13 分
		正确配置 sink			5 分
		创建好的 flume 组件组装为完整的 agent			8 分
	正确启动 Flume 采集数据	使用 flume-ng 命令启动配置的 agent			5 分

	验证数据处理是否成功	查看 logger sink 把文件数据打印到控制台的数据是否正确	10 分	
职业素养	专业素养	代码符合代码开发规范,命名规范,能做到见名知意;缩进统一,方便阅读;注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10 分	
总计		100 分		

12. 试题编号：2-2

(1) 任务描述

国内某 CDN 厂商，构建了全国性的服务网络，其为各大互联网厂商提供网络加速、内容分发等服务。其运营研发部为了方便收集开发日报、周报等数据。允许工作人员通过 tcp 协议发送指定格式的数据到某个端口。现委托某工程师进行技术调研，其初步选定使用 Flume 框架来完成数据的采集，其调研的重点是产品能够收集 tcp 协议发送的数据到 Flume 中。

本项目主要完成结合任务描述完成 Flume 的组件选择、画出相应的 agent 拓扑图、编写其对应的配置文件、启动 agent 处理数据并使用命令验证数据是否正确处理等功能。

任务一 选择合适的 Flume 组件（总共 10 分）

- (1) 根据项目描述，选择能够支持处理指定文件的 Flume source 组件(4分)；
- (2) 根据项目描述，选择能够进行快速测试的 Flume channel 组件（3分）；
- (3) 根据项目描述，选择能够实时展示测试数据的 Flume sink 组件（3分）；

任务二 画出 agent 的拓扑图（总共 10 分）

根据任务一选取的 Flume source 和 channel、sink 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 a1。（10分）

任务三 编写 agent 的配置文件（总共 45 分）

前期准备：

- 安装 netcat 工具：yum install -y nc
- 判断 44444 端口是否被占用：netstat -nlp | grep 44444
- 进入本地 flume 目录，新建 job 文件夹。

- (1) 根据任务描述和拓扑图，在 job 下创建 agent 的配置文件 a1_2_2.conf，确定所使用的 Flume source 组件的别名、Flume channel 组件的别名、Flume sink 组件的别名。（5分）

- (2) 编写前面创建的配置文件，定义好整个 agent 所使用的组件。（5分）

- (3) 编写 netcat tcp source 组件配置项：
 - a) 配置 source 组件的类型标识配置项(type)；（4分）

 - b) 配置 source 组件的监听的 IP 配置项(bind)为本机(localhost)；（4分）

 - c) 配置 source 组件监听的端口配置项(port)为 44444；（3分）

- (4) 编写 Memory channel 组件的配置项：
 - a) 配置 channel 组件的类型标识配置项(type)；（4分）

 - b) 配置 channel 组件容量大小配置项(capacity)为 1000；（3分）

 - c) 配置 channel 组件事务容量大小配置项(transactionCapacity)为 1000；（3分）

- (5) 编写 Logger sink 组件的配置项：
 - a) 配置 sink 组件的类型标识配置项（type）；（4分）

- (6) 将创建好的 flume 组件组装为完整的 agent
 - a) 配置 source 组件需要连接的 channel 的配置项（channels）（5分）

 - b) 配置 sink 组件需要连接的 channel 的配置项（channel）（5分）

任务四 使用 Flume 命令启动 agent 处理数据（总共 10 分）

- (1) 进入 flume 目录,使用 Flume 的安装目录下 bin 目录下的 flume-ng 脚本,

通过参数 agent 表示启动的是一个 agent，通过指定 flume 的配置文件所在目录(-c)、agent 的名称(-n)、agent 配置文件所在的目录(-f)来启动编写的 Flumeagent，并且通过-Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台。(5 分)

(2) 开启另一个会话，使用 netcat 工具向本机的 44444 端口发送消息 nc localhost 44444，发送内容为“hello word”；在 Flume 监听页面观察接收数据情况。(5 分)

任务五 验证数据是否正确处理（总共 5 分）

(1) logger sink 会将读取到指定文件的数据打印到控制台，将读取到的消息以及成功消费数据的截图存入到答案文件中。(5 分)

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-2 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图: XXXXXX, 并给出运行结果截图

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT、Hadoop 分布式系统	用以开发项目

测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	选择正确 Flume Source、channel、sink	选择的 Flumesource、channel、sink 是否符合要求	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	画出正确的拓扑图	根据选择的组件是否画出正确的 agent 图	10 分	
	编写 flume 配置文件	正确使用命令创建 agent 的配置文件 确定使用的 Flume 组件的别名 正确定义 agent 包含的组件	10 分	
		正确配置 source	11 分	
		正确配置 channel	10 分	
	正确配置 sink	4 分		

		创建好的 flume 组件组装为完整的 agent	10 分	
	正确启动 Flume 采集数据	使用 flume-ng 命令启动配置的 agent	5 分	
		向端口发送数据验证 flume 是否成功接收	5 分	
	验证数据处理是否成功	查看 logger sink 把文件数据打印到控制台的数据是否正确	5 分	
职业素养	专业素养	代码符合代码开发规范,命名规范,能做到见名知意;缩进统一,方便阅读;注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10 分	
总计		100 分		

13. 试题编号：2-3

(1) 项目描述

国内某 CDN 厂商，构建了全国性的服务网络，其为各大互联网厂商提供网络加速、内容分发等服务。客户在使用这些加速服务的时候，会产生服务日记，这些服务日记以文件的形式存在各个服务器上。CDN 厂商就是基于这些服务日记来计算带宽和流量，并以此作为收费依据。在完成新的日记数据进入消息队列外，还需要把原始的服务日记写入分布式存储系统 HDFS 中，以便数据重算和客户核对计费数据，因此需要将这些文件数据采集到 HDFS 中，作为备用数据源。

本项目主要完成结合任务描述完成 Flume 的组件选择、画出相应的 agent 拓扑图、编写其对应的配置文件、启动 agent 处理数据并使用命令验证数据是否正确处理等功能。

任务一 选择合适的 Flume 组件（总共 10 分）

- (1) 根据项目描述，选择能够支持处理指定文件的 Flume source 组件(4分)；
- (2) 根据项目描述，选择能够进行快速测试的 Flume channel 组件（3分）；
- (3) 根据项目描述，选择能够实时展示测试数据的 Flume sink 组件（3分）；

任务二 画出 agent 的拓扑图（总共 10 分）

根据任务一选取的 Flume source 和 channel、sink 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 a1。（10分）

任务三 编写 agent 的配置文件（总共 40 分）

前期准备：

- 进入本地 flume 目录，新建 testlogs 文件夹和 job 文件夹。在 testlogs 文件夹下新建文件 1.log、2.log、3.log、4.tmp，分别键入 test1、test2、test3、tmp4，保存。
- 启动 hadoop，进入 HDFS web 端。

- (1) 根据任务描述和拓扑图，在 job 文件夹下，创建 agent 的配置文件 a1_2

_3.conf, 确定所使用的 Flume source 组件的别名、Flume channel 组件的别名、Flume sink 组件的别名。(4分)

(2) 编写前面创建的配置文件, 定义好整个 agent 所使用的组件。(4分)

(3) 编写 spooldir source 组件配置项:

a) 配置 source 组件的类型标识配置项 (type); (2分)

b) 配置 source 组件监听的目录配置项 (spoolDir) 为 testlogs 目录; (2分)

c) 配置 source 组件处理完成文件后的后缀名配置项 (fileSuffix) 为.COMPLETED; (2分)

d) 配置包含文件头选项 (fileHeader) 为 true; (2分)

e) 配置 source 组件匹配出不需要的处理文件名正则表达式配置项 (ignorePattern) 为([^\.]*\.tmp); (2分)

(4) 编写 Memory channel 组件的配置项:

a) 配置 channel 组件的类型标识配置项 (type) 为 memory; (2分)

b) 配置 channel 组件容量大小配置项 (capacity) 为 1000; (2分)

c) 配置 channel 组件事务容量大小配置项 (transactionCapacity) 为 100; (2分)

(5) 编写 HDFS sink 组件的配置项:

a) 配置 sink 组件的类型标识配置项 (type); (2分)

- b) 配置 sink 组件的数据，写入到 HDFS 的/flume/upload/%Y-%m-%d 路径；
(2 分)
 - c) 配置 sink 组件将数据写入 HDFS 完成后文件前缀配置项(hdfs.filePrefix)为 upload-； (2 分)
 - d) 配置 sink 组件每隔多长时间完成一次文件写入的配置项(hdfs.rollInterval)为 10； (2 分)
 - e) 配置 sink 组件每批次处理数据量的批次大小(hdfs.batchSize)为 100；
(2 分)
 - f) 配置 sink 组件使用本地时间(hdfs.useLocalTimeStamp)为 true； (2 分)
- (6) 将创建好的 Flume 组件组装为完整的 agent
- a) 配置 source 组件需要连接的 channel 的配置项(channels)； (2 分)
 - b) 配置 sink 组件需要连接的 channel 的配置项(channel)； (2 分)

任务四 使用 Flume 命令启动 agent 处理数据 (总共 5 分)

- (1) 进入 flume 目录,使用 Flume 的安装目录下 bin 目录下的 flume-ng 脚本,通过参数 agent 表示启动的是一个 agent,通过指定 flume 的配置文件所在目录(-c)、agent 的名称(-n)、agent 配置文件所在的目录(-f)来启动编写的 Flumeagent,并且通过-Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台。(5 分)

任务五 验证数据是否正确处理 (总共 15 分)

- (1) 查看 testlogs 下的文件后缀名是否添加. COMPLETE; (5 分)
- (2) 使用 hdfs 命令或者是 web 端查看文件是否成功写入到 hdfs; (5 分)
- (3) 查看 testlogs 下的文件内容, 并使用 hdfs 命令或者 web 端查看 flume 写入的内容是否正确; (5 分)

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹, 考生文件夹的命名规则: 考生学校+考生号+考生姓名, 示例: 永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-3 答案.docx”文件, 文件内容格式为“任务号+题号+命令详情+答案”, 示例如下:

任务一 (1) 命令和截图: XXXXXX, 并给出运行结果截图

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计, 每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT、Hadoop 分布式系统	用以开发项目
测评专家	现场测评专家: 在本行业具有 3 年以上的从业经验 (工程师及以上职称) 或从事本专业具有 5 年以上的教学经验 (副高及以上职称), 或具有软件设计师、系统分析师、数据库设计师资格证书 (2 人/场)。		测评专家满足任一条件
	结果测评专家: 在本行业具有 3 年以上的从业经验 (工程师及以上职称) 或从事本专业具有 5 年以上的教学经验 (副高及以上职称), 或具有软件设计师、系统分析		

师、数据库设计师资格证书（2人/场）。	
---------------------	--

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注	
工作任务	选择正确 Flume Source、channel、sink	选择的 Flumesource、channel、sink 是否符合要求	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。	
	画出正确的拓扑图	根据选择的组件是否画出正确的 agent 图	10 分		
	编写 flume 配置文件	正确使用命令创建 agent 的配置文件 确定使用的 Flume 组件的别名 正确定义 agent 包含的组件			8 分
		正确配置 source			10 分
		正确配置 channel			6 分
		正确配置 sink			12 分
		创建好的 flume 组件组装为完整的 agent			4 分
	正确启动 Flume 采集数据	使用 flume-ng 命令启动配置的 agent			5 分

	验证数据处理是否成功	查看 HDFS sink 把文件是否监控导入	15 分	
职业素养	专业素养	代码符合代码开发规范,命名规范,能做到见名知意;缩进统一,方便阅读;注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10 分	
总计		100 分		

项目 2：基于 kafka 的消息队列数据采集

14. 试题编号：2-4

(1) 任务描述

随着中国汽车市场的飞速发展，城市汽车保有量也呈现高速增长，城市交通压力也越来越大。为了更好的疏导城市交通，借助于基于神经网络的深度学习技术，对城市交通摄像头的视频数据进行处理，生成车辆结构化数据并以文件的形式进行保存，格式为 txt，并且需要把生成的新文件数据实时采集到消息队列中 (Kafka) 去。

本项目主要完成结合任务描述完成 Flume 的组件选择、画出相应的 agent 拓扑图、编写其对应的配置文件、创建 Kafka Topic 用来存储数据、启动 agent 处理数据并使用命令验证数据是否正确处理等功能。

任务一 选择合适的 Flume 组件（总共 10 分）

- (1) 根据项目描述,选择能够支持处理指定文件的 Flume source 组件(5分);
- (2) 根据项目描述,选择能够进行快速测试的 Flume channel 组件(5分);

任务二 画出 agent 的拓扑图（总共 10 分）

根据任务一选取的 Flume source 和 channel、sink 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 a1。（10分）

任务三 编写 agent 的配置文件（总共 40 分）

前期准备：

- 进入本地 flume 目录，新建 logs 文件夹和 job 文件夹。在 logs 文件夹下新建文件 1.log，键入 test 后保存。
- 启动 zookeeper。

- (1) 根据任务描述和拓扑图，在 job 目录下，创建 agent 的配置文件 a1_2_4.conf, 确定所使用的 Flume source 组件的别名、Flume channel 组件的别名、Flume sink 组件的别名。（5分）

- (2) 编写前面创建的配置文件，定义好整个 agent 所使用的组件。（5分）
- (3) 编写 exec source 组件配置项：
- c) 配置 source 组件的类型标识配置项 (type)；（5分）
 - d) 配置 source 组件监听的 Unix 命令或者脚本 (command)，监听 1.log；（5分）
- (4) 编写 Kafka channel 组件的配置项：
- 配置 channel 组件的类型标识配置项 (type) 为 org.apache.flume.channel.kafka.KafkaChannel；（5分）
- a) 配置 channel 组件服务器 IP 和端口配置项 (kafka.bootstrap.servers) 为 localhost:9092；（5分）
 - b) 配置 channel 组件数据写入的 Topic 名称配置项 (kafka.topic) 为 words；（5分）
- (5) 将创建好的 flume 组件组装为完整的 agent：
- a) 配置 source 组件需要连接的 channel 的配置项 (channels)（5分）

任务四 使用命令根据 agent 的配置文件，创建 KafkaTopic（总共 10 分）

- (1) 启动 kafka 服务端，创建主题 (topic) words。使用 Kafka 的 bin 目录下的 kafka-topics.sh 脚本，使用参数 (--create) 表示创建 topic，指定连接的 Kafka 服务器 IP 和端口 (--bootstrap-server)。创建一个备份数 (--replication-factor) 为 1，分区数 (--partitions) 为 1 的 topic。该 topic 的名字为 agent 配置文件中 channel 模块中配置的 topic 名称 (--topic)。（5分）

任务五 使用 Flume 命令启动 agent 处理数据（总共 5 分）

- (1) 进入 flume 目录，使用 Flume 的安装目录下的 bin 目录下的 flume-ng 脚本，通过参数 agent 表示启动一个完整 agent，参数通过指定 flume 的配置文件所在目录(-c)、agent 的名称(-n)、agent 配置文件所在的目录(-f)来启动编写的 Flumeagent，并且通过-Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台。将 agent 的启动命令和运行界面截图执行结果存放到答案文件中。（10 分）

任务六 验证数据是否正确处理（总共 5 分）

- (1) 新建一个会话，启动 kafka 消费端，使用 Kafka 的 bin 目录下的 kafka-console-consumer.sh 命令，通过指定连接的 Kafka 服务器 IP 和端口(--bootstrap-server)、需要读取的 Topic(--topic)以及从什么位置开始消费(--from-beginning)验证数据是否写入指定的 topic 中。将读取消息的命令以及成功消费数据的截图存入到答案文件中。（5 分）

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-4 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图: XXXXXX, 并给出运行结果截图

3) 将配置文件.conf 保存到考生文件夹。

4) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Centos7 或更高版本	用于程序设计，每人一台。

	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT	用以连接服务器
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

（4）考核时量

考核时间为 90 分钟

（5）评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	选择正 Flume Source、channel	选择的 Flumesource 和 channel 是否符合要求	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	画出正 agent 图	根据选择的组件是否画出正确的 agent 图	10 分	
	编写 flume 配置文件	正确使用命令创建 agent 的配置文件 确定使用的 Flume 组件的别名 正确定义 agent 包含的组件 配置 source 组件的类型标识配置项 配置 source 组件监听的目录配置项 配置 source 组件处理完成文件后的后缀名配置项 配置 source 组件匹配出需要的处理文件名正则表达式配	10 分	

		置项 配置 source 组件匹配出不需要的处理文件名正则表达式 配置项		
		配置 channel 组件的类型标识配置项 配置 channel 组件服务器 IP 和端口配置项 配置 channel 组件数据写入的 Topic 名称配置项	15 分	
		创建好的 flume 组件组装为完整的 agent	5 分	
	创建 agent 配置的 Topic	正确使用 Kafka-topic.sh 命令创建备份数为 1、分区为 1 的 Topic	5 分	
	正确启动 Flume 采集数据	使用 flume-ng 命令启动配置的 agent	10 分	
	验证数据处理是否成功	使用 kafka-console-consumer.sh 命令验证数据是否正确处理	5 分	
职业素养	专业素养	代码符合代码开发规范,命名规范,能做到见名知意;缩进统一,方便阅读;注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10 分	
总计			100 分	

15. 试题编号：2-5

(1) 任务描述

国内某 CDN 厂商，构建了全国性的服务网络，其为各大互联网厂商提供网络加速、内容分发等服务。客户在使用这些加速服务的时候，会产生服务日记，这些服务日记以文件的形式存在各个服务器上。CDN 厂商就是基于这些服务日记来计算带宽和流量，并以此作为收费依据。因此需要构建分布式的文件收集系统，将这些文件数据采集到消息队列中，作为后续流式计算的数据源。

本项目主要完成结合任务描述完成 Flume 的组件选择、画出相应的 agent 拓扑图、编写其对应的配置文件、创建 KafkaTopic 用来存储数据、启动 agent 处理数据并使用命令验证数据是否正确处理等功能。

任务一 选择合适的 Flume 组件（总共 10 分）

- (1) 根据项目描述，选择能够支持处理指定文件的 Flume source 组件(4分)；
- (2) 根据项目描述，选择能够进行快速测试的 Flume channel 组件（3分）；
- (3) 根据项目描述，选择能够实时展示测试数据的 Flume sink 组件（3分）；

任务二 画出 agent 的拓扑图（总共 10 分）

根据任务一选取的 Flume source 和 channel、sink 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 a1。将画出的拓扑图截图并命名为“试题编号 2-5 拓扑图”，将其存放到考生文件夹中。（10 分）

任务三 编写 agent 的配置文件（总共 40 分）

前期准备：

- 进入本地 flume 目录，新建 testlogs 文件夹和 job 文件夹。在 testlogs 文件夹下新建文件 1.log、2.log、3.log、4.tmp，分别键入 test1、test2、test3、tmp4，保存。
 - 启动 zookeeper。
- (1) 根据任务描述和拓扑图，在 job 文件夹下，创建 agent 的配置文件 a1_2

_5.conf, 确定所使用的 Flume source 组件的别名、Flume channel 组件的别名、Flume sink 组件的别名。(5 分)

(2) 编写前面创建的配置文件, 定义好整个 agent 所使用的组件。(5 分)

(3) 编写 spooldir source 组件配置项:

a) 配置 source 组件的类型标识配置项 (type); (4 分)

b) 配置 source 组件监听的目录配置项 (spoolDir) 为 testlogs 目录; (2 分)

c) 配置 source 组件处理完成文件后的后缀名配置项 (fileSuffix) 为.COMPLETED; (2 分)

d) 配置包含文件头选项 (fileHeader) 为 true; (2 分)

e) 配置 source 组件匹配出不需要的处理文件名正则表达式配置项 (ignorePattern) 为([^\.]*\.tmp); (2 分)

(4) 编写 Memory channel 组件的配置项:

a) 配置 channel 组件的类型标识配置项 (type) 为 memory; (4 分)

b) 配置 channel 组件容量大小配置项 (capacity) 为 1000; (2 分)

c) 配置 channel 组件事务容量大小配置项 (transactionCapacity) 为 100; (2 分)

(5) 编写 Kafka sink 组件的配置项:

a) 配置 sink 组件的类型标识配置项 (type) 为 org.apache.flume.sink.k

afka.KafkaSink; (2分)

b) 配置 sink 组件的连接 Kafka 集群的 IP 和端口(kafka.bootstrap.servers)为 localhost:9092; (2分)

c) 配置 sink 组件将数据读出写入的 Topic 配置项(kafka.topic)为 words (2分)

(6) 将创建好的 Flume 组件组装为完整的 agent

a) 配置 source 组件需要连接的 channel 的配置项 (channels) (2分)

b) 配置 sink 组件需要连接的 channel 的配置项 (channel) (2分)

任务四 使用命令根据 agent 的配置文件，创建 KafkaTopic (总共 5 分)

(1) 使用 Kafka 的 bin 目录下的 kafka-topics.sh 脚本，使用参数 (--create) 表示创建 topic，指定连接的 Kafka 服务器 IP 和端口(--bootstrap-server)。创建一个备份数(--replication-factor) 为 1，分区数 (--partitions) 为 1 的 topic。该 topic 的名字为 agent 配置文件中 channel 模块中配置的 topic 名称 (--topic)。将创建 topic 的命令以及创建成功的标识截图存放到答案文件中。(5分)

任务五 使用 Flume 命令启动 agent 处理数据 (总共 10 分)

(1) 使用 Flume 的安装目录下 bin 目录下的 flume-ng 脚本，通过参数 agent 表示启动一个完整 agent，通过指定 flume 的配置文件所在目录(-c)、agent 的名称(-n)、agent 配置文件所在的目录(-f)来启动编写的 Flumeagent，并且通过-Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台。将 agent 的启动命令和运行界面截图执行结果存放到答案文件中。(10分)

任务六 验证数据是否正确处理（总共 5 分）

- (1) 使用 Kafka 的 bin 目录下的 kafka-console-consumer.sh 命令，通过指定连接的 Kafka 服务器 IP 和端口(--bootstrap-server)、需要读取的 Topic(--topic)以及从什么位置开始消费(--from-beginning) 验证车辆数据是否写入到相应的 topic 中。将读取消息的命令以及成功消费数据的截图存入到答案文件中。（5 分）

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-5 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 将配置文件.conf 保存到考生文件夹。

4) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT	用以连接服务器
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件

结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。
--

（4）考核时量

考核时间为 90 分钟

（5）评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	选择正Flume Source、channel、sink	选择的 Flumesource 和 channel、sink 是否符合要求	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	画出正 agent 图	根据选择的组件是否画出正确的 agent 图	10 分	
	编写 flume 配置文件	正确使用命令创建 agent 的配置文件 确定使用的 Flume 组件的别名 正确定义 agent 包含的组件	10 分	
		配置 source 组件的类型标识配置项 配置 source 组件监听的目录配置项 配置 source 组件处理完成文件后的后缀名配置项 配置 source 组件匹配出需要的处理文件名正则表达式配置项 配置 source 组件匹配出不需要的处理文件名正则表达式配置项	12 分	
		配置 channel 组件的类型标识配置项； 配置 channel 组件数据缓存目录配置项； 配置 channel 组件元数据 check point 缓存目录配置项； 配置 channel 组件容量大小配	8 分	

		置项; 配置 channel 组件事务容量大小配置项;		
		配置 sink 组件的类型标识配置项; 配置 sink 组件的连接 Kafka 集群的 IP 和端口; 配置 sink 组件将数据读出写入的 Topic 配置项	5 分	
		创建好的 flume 组件组装为完整的 agent	5 分	
	创建 agent 配置的 Topic	正确使用 Kafka-topic.sh 命令 创建备份数为 1、分区为 1 的 Topic 以存放车辆数据	5 分	
		使用 flume-ng 命令启动配置的 agent	10 分	
	验证数据处理是否成功	使用 kafka-console-consumer.sh 命令验证车辆数据是否正确处理	5 分	
职业素养	专业素养	代码符合代码开发规范, 命名规范, 能做到见名知意; 缩进统一, 方便阅读; 注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明, 遵守考场纪律, 按顺序进出考场。	0-10 分	
总计		100 分		

项目 3：基于 HBase 的非结构化数据存储

16. 试题编号：2-6

(1) 任务描述

为调查某一地区学生计算机学习能力，需收集学生的计算机学习情况，包括 Python、Java 和 math 成绩，因为该地区学生数量大，且很多学生的数据收集不齐全，因此不能采用结构化数据存储，从而导致大量的空值占用存储空间。所以，需要采用 HBase 数据库，针对列存储，以下是关于该项目 HBase 存储非结构化数据的操作。

任务一 命名空间与建表操作（每题 3 分，总共 30 分）

- (1) 根据项目描述，启动 hadoop、hbase，查看进程，验证是否成功启动 hbase、hadoop 进程。
- (2) 创建一个命名空间 myns，同时设置属性 company 为 hnzj
- (3) 列出 HBase 中所有的命名空间
- (4) 在命名空间 myns 下创建如下表格，列族 info 的版本数为 3 个，列族 course 的版本数为 3 个。并验证是否成功创建。

表 : score				
info		course		

- (5) 列出 hbase 命令空间 myns 下的所有的表
- (6) 获取上表的详细信息
- (7) 给上表增加一个列族，列族名为 department，版本数为 3 个

- (8) 修改列族 info 的版本数为 2
- (9) 查看表 score 当前的列族信息
- (10) 列出命名空间 myns 下的所有表

任务二 数据操作（每个单元格的数据 1 分，总共 30 分）

- (1) 向任务一的表 score 中插入数据，如下所示：

表 : score						
行键	info		course			department
	name	grade	java	python	math	id
1001	bob		90	95	66	A
1002	alice	1	56		45	B
3001	clerk		88	87	76	
2001		2		89		C
2002	jerry		96	90	89	
3002	luna	3	67	50		C
1003	lucy	1		88	76	

任务三 数据查询操作（每题 4 分，总共 20 分）

- (1) 查询全表数据
- (2) 统计行数
- (3) 查询表的前 3 行数据

(4) 查询 1003 号学生的 java、python 成绩信息

(5) 使用行过滤器，扫描 2002 号学生的所有数据

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-6 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	命名空间、创建数据表操作	命名空间、创建数据表操作正确	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	创建数据表操作	插入数据操作正确	30 分	
	数据查询操作	数据查询操作正确	20 分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分		

17. 试题编号：2-7

(1) 任务描述

为调查某一地区学生计算机学习能力，需收集学生的计算机学习情况，包括 Python、Java 和 math 成绩，因为该地区学生数量大，且很多学生的数据收集不齐全，因此不能采用结构化数据存储，从而导致大量的空值占用存储空间。所以，需要采用 HBase 数据库，针对列存储，以下是关于该项目 HBase 存储非结构化数据的操作。

任务一 建表与插入数据操作（总共 40 分）

- (1) 在默认命名空间下，创建如下表格，列族 info 的版本数为 3 个，列族 course 的版本数为 2 个。并验证是否成功创建。（10 分）

表：score						
info		course			department	

- (2) 向任务一的表 score 中插入数据（每个单元格的数据 1 分，总共 30 分），如下所示：

表：score						
行键	info		course			department
	name	grade	java	python	math	id
1001	bob		90	95	66	A
1002	alice	1	56		45	B
3001	clerk		88	87	76	
2001		2		89		C
2002	jerry		96	90	89	

3002	luna	3	67	50		C
1003	lucy	1		88	76	

任务二 数据查询操作（每题 4 分，总共 20 分）

- (1) 查询 1002 号学生值大于 60 的数据
- (2) 查询学生 luna 模糊匹配 6 的数据
- (3) 查询表中包含 'lu' 的数据
- (4) 查询表中大于等于 90 分的成绩
- (5) 查询表中大于 80，小于等于 90 分的成绩

任务三 过滤器操作（每题 5 分，总共 20 分）

- (1) 使用行过滤器 RowFilter，扫描 2002 号学生的所有数据
- (2) 使用列族过滤器 FamilyFilter，扫描 course 列族的所有数据
- (3) 使用列名过滤器 QualifierFilter，扫描 java 这一列数据
- (4) 使用单列值过滤器 SingleColumnValueFilter，扫描 lucy 的数据

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-7 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	建表与插入数据操作	建表与插入数据操作正确	40 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、
	数据查询操作	数据查询正确	20 分	
	过滤器操作	过滤器操作正确	20 分	

职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	造成恶劣影响的本项目记0分。
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分		

18. 试题编号：2-8

(1) 任务描述

为调查某一地区学生计算机学习能力，需收集学生的计算机学习情况，包括 Python、Java 和 math 成绩，因为该地区学生数量大，且很多学生的数据收集不齐全，因此不能采用结构化数据存储，从而导致大量的空值占用存储空间。所以，需要采用 HBase 数据库，针对列存储，以下是关于该项目 HBase 存储非结构化数据的操作。

任务一 命名空间与建表操作（每题 3 分，总共 15 分）

- (1) 根据项目描述，选启动 hadoop、hbase，查看进程，验证是否成功启动 hbase、hadoop 进程。
- (2) 创建一个命名空间 myns，同时设置属性 company 为 hnzj
- (3) 在默认命名空间下，创建如下表格，列族 info 的版本数为 3 个，列族 course 的版本数为 3 个。并验证是否成功创建。

表 : score				
info		course		

- (4) 列出 hbase 命令空间下的所有的表
- (5) 获取上表的详细信息

任务二 数据插入操作（每个单元格的数据 1 分，总共 20 分）

- (1) 向任务一的表 score 中插入数据，如下所示：

表 : score		
行键	info	course

	name	grade	java	python	math
1001	bob		90	95	66
1002	alice	1	56		45
3001	clerk		88	87	76
2002	jerry		96	90	89
3002	luna	3	67	50	

任务三 数据与过滤器操作（每题 5 分，总共 45 分）

- (1) 扫描全表数据
- (2) 修改 1002 号学生的 java 成绩为 60，math 成绩为 50
- (3) 查询每个人的姓名和所有成绩
- (4) 查询 math 成绩及格的数据
- (5) 从行键 3001 开始扫描，只扫描 2 行数据
- (6) 使用列族过滤器 FamilyFilter，扫描 course 列族的所有数据
- (7) 使用单列排除过滤器 SingleColumnValueExcludeFilter，扫描 clerk 的其他数据
- (8) 删除 1001 号学生的 math 成绩，并进行验证
- (9) 清空表中数据，并进行验证

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 2-8 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一（1）命令和截图： XXXXXX， 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	命名空间与建表操作	命名空间与建表操作正确	15 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。
	数据插入操作	数据插入正确	20 分	

	数据和过滤器操作	数据和过滤器操作正确	45分	2、严重违反考场纪律、造成恶劣影响的本项目记0分。
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分		

模块三 数据清洗与分析

项目 1：基于 pandas 的数据清洗

19. 试题编号：3-1

(1) 任务描述

“居”是民生的重要组成部分，也是百姓幸福生活的重要保障。为了增进民生福祉，提高人民生活品质，对房地产市场进行精准调控决策，现需分析和统计某地区房屋销售数据。该地区房屋销售数据主要存放了房屋售出时间、地区邮编、房屋价格、房屋类型和配套房间数 5 个特征，部分数据如下表所示，其中房屋类型有普通住宅 (house) 和单身公寓 (unit) 两种。探索数据的基本信息，通过索引操作查询到房屋类型为单身公寓的数据，同时观察数据的整体分布并发现数据间的关联。注意，地区邮编特征已完成脱敏处理，因此只存在 4 位数。

房屋出售时间	地区邮编	房屋价格	房屋类型	配套房间数
2010/1/4 0:00	2615	435000	house	3
2010/1/5 0:00	2904	712000	house	4
2010/1/6 0:00	2617	435000	house	4
2010/1/6 0:00	2606	1350000	house	5
2010/1/7 0:00	2905	612500	house	4

任务一：读取数据与描述数据情况（每题 5 分，总共 30 分）

- (1) 读取“某地区房屋销售数据.csv”文件，命名为 house_info
- (2) 获取数据的行索引、列名、每列数据类型、数据元素值
- (3) 使用 numpy 的描述性统计函数，统计房屋价格的最大值、最小值、均值
- (4) 使用 numpy 的描述性统计函数，描述房屋价格数据之间的离散程度(std)
- (5) 使用 pandas 的描述性统计函数，统计配套房间数的最大值、最小值、均值

- (6) 使用 pandas 的描述性统计函数，统计配套房间数的离散程度(std)

任务二：数据统计（每题 5 分，总共 25 分）

- (1) 统计普通住宅(house)和单身公寓(unit)的频数(value_counts())
- (2) 对房屋价格这一列，实现数值型特征的描述性统计(describe)，并对这些特征结果进行描述。
- (3) 将“房屋出售时间”的月份提取出来，作为新列插入，新列的名称为“房屋出售月份”；进行验证。
- (4) 按“房屋类型”进行分组，求每组的大小
- (5) 按“房屋类型”进行分组，求每组“房屋价格”的均值

任务三：数据清洗（每题 5 分，总共 25 分）

- (1) 检测每列的空值数量
- (2) 在 house_info 上，删除有空值的行，不在原数据上操作；并进行验证
- (3) 在 house_info 上，根据“房屋类型”，利用 drop_duplicates() 方法去重，保留重复的第一个数据，不在原数据上操作。
- (4) 在 house_info 上，删除有空值的列，不在原数据上操作；并进行验证。
- (5) 在 house_info 上，根据“房屋类型”，利用 drop_duplicates() 方法去重，不保留重复的数据，不在原数据上操作

(2) 提交要求

- 1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的

命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-1 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX, 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT	用以连接服务器
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容	评分标准	备注
------	------	----

工作任务	导入数据与数据描述	导入数据与数据描述正确	30分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	数据统计操作	数据统计正确	25分	
	数据清洗操作	数据清洗操作正确	25分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分		

20. 试题编号：3-2

(1) 任务描述

我国始终把保障人民健康放在优先发展的战略位置。“上医治未病”，建立疾病预防控制体系有利于从源头上预防和控制重大疾病。某医院为了早期监测预警患者的中风风险，对现有中风患者的基础信息和体检数据 (healthcare-dataset-stroke.xlsx) 进行分析，其部分数据如下表所示。

编号	性别	高血压	是否结婚	工作类型	居住类型	体重指数	吸烟史	中风
9046	男	否	是	私人	城市	36.6	以前吸烟	是
51676	女	否	是	私营企业	农村	NaN	从不吸烟	是
31112	男	否	是	私人	农村	32.5	从不吸烟	是
60182	女	否	是	私人	城市	34.4	抽烟	是
1665	女	是	是	私营企业	农村	24.0	从不吸烟	是

经观察发现患者基础信息中缺少中风患者的年龄和平均血糖的信息，然而在年龄和平均血糖数据 (healthcare-dataset-age_abs.xlsx) 中存放了分析所需的中风患者的年龄和平均血糖信息，其部分数据如下表所示。

编号	年龄	平均血糖
9046	67.0	228.69
51676	61.0	202.21
31112	80.0	105.92
60182	49.0	171.23
1665	79.0	174.12

现需要对患者的年龄、平均血糖数据与患者基础信息和体检数据进行合并，并进一步分析。

任务一：读取数据与描述数据情况（每题 5 分，总共 30 分）

- (1) 读取中风患者的基础信息和体检数据 (healthcare-dataset-stroke.xlsx) 文件，命名为 info1
- (2) 读取中风患者的年龄和平均血糖数据 (healthcare-dataset-age_abs.xlsx) 文件，命名为 info2

- (3) 查看 info1 数据的基本属性 values、index、columns、dtypes
- (4) 查看 info2 数据的基本属性 values、index、columns、dtypes
- (5) 查看 info1 的前 5 行数据
- (6) 查看 info2 的前 2 列数据

任务二：数据统计（每题 5 分，总共 25 分）

- (1) 以编号为主键，对两个数据进行外连接合并，合并后的数据命名为 merge_info
- (2) 读取 merge_info 的前 20 行数据，并对进行“工作类型”哑变量处理 (get_dummies())
- (3) 在 merge_info 上，统计每个特征的缺失值数量
- (4) 在 merge_info 上，计算“年龄”、“体重指数”、“平均血糖”的相似度矩阵(corr)。
- (5) 使用等宽法，对“年龄”数据进行离散化处理(cut)，处理成 4 个年龄段的数据

任务三：数据清洗（每题 5 分，总共 25 分）

- (1) 在 merge_info 上，根据“工作类型”，利用 drop_duplicates() 方法去重，保留重复的第一个数据，不在原数据上操作。
- (2) 在 merge_info 上，使用 drop_duplicates() 方法对“是否结婚”、“工作类型”去重，不保留重复元素，不在原数据上操作。
- (3) 在 merge_info 上，将“体重指数”的空值，用“体重指数”的平均值替换，并进行验证。

(4) 在 merge_info 上，删除有空值的行，不在原数据上操作；并进行验证。

(5) 在 merge_info 上，删除有空值的列，不在原数据上操作；并进行验证。

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-2 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT	用以连接服务器
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	导入数据与数据描述	导入数据与数据描述正确	20 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	数据统计操作	数据统计正确	25 分	
	数据清洗操作	数据清洗操作正确	35 分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分		

21. 试题编号：3-3

(1) 任务描述

某公司人事工作人员为了对来聘人员信息并行分析，以聘用适合计算机岗位的人员，调用了计算机岗位来聘人员信息表(hr_job.csv)，其部分数据如表 4-17 所示，数据字段包括招聘人员 ID、性别、相关经验、教育水平和工作次数等信息。其部分数据如下表所示：

应聘人员ID	性别	相关经验	教育水平	工作次数
11561	NaN	无	大学	5.0
33241	NaN	无	大学	0.0
21651	NaN	有	大学	11.0
28806	男	有	高中	5.0
402	男	有	大学	13.0

经观察发现，数据存在缺失值等异常数据，因此需要对数据进行预处理。

任务一：读取数据与描述数据情况（每题 5 分，总共 30 分）

- (1) 读取计算机岗位来聘人员信息表(hr_job.csv)文件，命名为 hr_info
- (2) 读取 hr_job.csv 的前 5 行数据
- (3) 读取 hr_job.csv 的后 3 行数据
- (4) 使用 loc 读取 hr_job.csv 的列，行名称为 1、3
- (5) 使用 iloc 读取 hr_job.csv 的列，列位置为 2、4
- (6) 查看 hr_info 数据的基本属性 values、index、columns、dtypes

任务二：数据统计（每题 5 分，总共 25 分）

- (1) 利用 set 的特征去重，统计“教育水平”去重后的种类、种类出现的个数

(2) 对工作次数这一列，实现数值型特征的描述性统计（describe），并对这些特征结果进行描述。

(3) 统计最多工作次数

(4) 统计每列的空值

(5) 统计所有应聘人员工作次数的平均值

任务三：数据清洗（每题 6 分，总共 25 分）

(1) 在 hr_info 上，删除有空值的行，不在原数据上操作；并进行验证

(2) 在 hr_info 上，删除有空值的列，不在原数据上操作；并进行验证

(3) 将数值型缺失值填补（fillna）为其对应特征的均值，并进行验证

(4) 对所有的分类数据进行哑变量处理（get_dummies()）

(5) 对工作次数离散化处理（cut）成 3 类

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-3 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT	用以连接服务器
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	导入数据与数据描述	导入数据与数据描述正确	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。
	数据统计操作	数据统计正确	25 分	

	数据清洗操作	数据清洗操作正确	25 分	2、严重违反考场纪律、造成恶劣影响的本项目记0分。
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100 分		

22. 试题编号：3-4

(1) 任务描述

“居”是民生的重要组成部分，也是百姓幸福生活的重要保障。为了增进民生福祉，提高人民生活品质，对房地产市场进行精准调控决策，现需分析和统计某地区房屋销售数据。该地区房屋销售数据主要存放了房屋售出时间、地区邮编、房屋价格、房屋类型和配套房间数 5 个特征，部分数据如下表所示，其中房屋类型有普通住宅(house)和单身公寓(unit)两种。探索数据的基本信息，通过索引操作查询到房屋类型为单身公寓的数据，同时观察数据的整体分布并发现数据间的关联。注意，地区邮编特征已完成脱敏处理，因此只存在 4 位数。

房屋出售时间	地区邮编	房屋价格	房屋类型	配套房间数
2010/1/4 0:00	2615	435000	house	3
2010/1/5 0:00	2904	712000	house	4
2010/1/6 0:00	2617	435000	house	4
2010/1/6 0:00	2606	1350000	house	5
2010/1/7 0:00	2905	612500	house	4

任务一：读取数据与描述数据情况（每题 5 分，总共 30 分）

- (1) 读取“某地区房屋销售数据.csv”文件，命名为 house_info
- (2) 获取数据的行索引、列名、每列数据类型、数据元素值
- (3) 使用 tail() 方法读取 house_info 的后 3 行数据
- (4) 使用 head() 方法读取 house_info 的前 3 行数据
- (5) 使用 pandas 的描述性统计函数(describe)，统计配套房间数的最大值、最小值、均值
- (6) 使用 pandas 的描述性统计函数，统计配套房间数的离散程度(std)

任务二：数据统计（每题 5 分，总共 25 分）

- (1) 使用聚合方法，求出“房屋价格”、“出售数量”的总和和均值
- (2) 新增列“出售数量”，插入数据 1；进行验证
- (3) 将“房屋出售时间”的月份提取出来，作为新列插入，新列的名称为“房屋出售月份”；进行验证
- (4) 按“房屋类型”进行分组，求每组的大小
- (5) 更改“房屋价格”列名为“房屋价格（万元）”

任务三：数据清洗（每题 5 分，总共 25 分）

- (1) 检测每列的空值数量
- (2) 在 house_info 上，删除有空值的行，不在原数据上操作；并进行验证
- (3) 在 house_info 上，删除有空值的列，不在原数据上操作；并进行验证
- (4) 筛选出房屋类型为 house 的数据，并进行验证
- (5) 筛选出房屋类型为 unit 的数据，并进行验证

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-4 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX, 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	XShell、SecureCRT	用以连接服务器
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	导入数据与数据描述	导入数据与数据描述正确	30 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。
	数据统计操作	数据统计正确	25 分	

	数据清洗操作	数据清洗操作正确	25 分	2、严重违反考场纪律、造成恶劣影响的本项目记0分。
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100 分		

项目 2：基于 kettle 的数据清洗

23. 试题编号：3-5

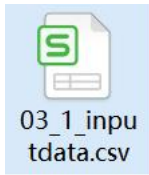
(1) 任务描述

无人售货机在我们日常生活中已经较为普遍，它是一种根据投币或者扫码支付的自动付货的机器，某无人售货机公司采集无人售货机订单信息数据（03_1_inputdata.csv）的部分关键字段说明如下表 1 所示，源数据文件类型为分隔符的数据类型，为了了解客户订单的状态，要求利用 Kettle 软件对该数据进行过滤、字段选择、排序等清洗操作。

表 1 无人售货机订单信息数据字段说明

字段名称	位置	说明	示例
createdtime	1	订单生成的时间	2018/11/21 18:46:13
customerid	2	客户 ID	204524
customermobile	3	客户手机号码	18660951385
totalprice	4	订单总额	4.5
paytotalprice	5	订单实际支付金额	4.5
discounttotalprice	6	订单优惠金额	0
status	7	订单状态	SUCCESS
source	8	订单来源	ALIPAY_SHOPPING
boxid	9	售货机 ID	73216297342
ordernum	10	订单号	272074322789605000
payexceptiontype	11	订单异常菜单列表	COMMON
payedtime	12	支付时间	2018/11/21 18:46:14
productname	13	商品名称	康师傅经典红烧牛肉面
amount	14	商品数量	1
costprice	15	商品成本价	3.58
saleprice	16	商品销售价	4.5
productpaytotalprice	17	商品实际支付总金额	4.5
producttotalprice	19	商品支付总金额	4.5

源数据文档名称：03_1_inputdata.csv



任务一：数据的导入（15分）

1. 在 Kettle 软件中新建名为 03-1 的转换文件（.ktr），选择输入类别中的 CSV 文件输入步骤。（7分）

2. 配置 CSV 文件输入步骤中的相关参数，将 03_1_inputdata.csv 中的源数据全部导入 kettle 平台中（注：customermobile 列需改类型为 String）。（8分）

任务二：数据的清洗（50分）

1. 对获取的数据进行过滤，选择流程类别中的过滤记录步骤，建立 CSV 文件输入步骤到过滤记录步骤之间的连接，设置参数，过滤掉客户 ID 为空和支付不成功的订单数据。

2. 对获取的数据进行抽取，选择转换类别中的字段选择步骤，建立过滤记录步骤到字段选择步骤之间的连接，只保留需要的 customerid、customermobile、ordernum 和 paytotalprice 字段，并且分别改为客户 ID、客户 TEL、订单数量和支付总价。

3. 对已经过滤和抽取的数据进行聚合统计，选择转换类别中的排序记录步骤，建立字段选择步骤到排序记录步骤之间的连接，设置按照“客户 ID”字段升序排序。

4. 对客户的订单数和商品实际支付金额等字段进行分组聚合，统计各个客户的订单。选择统计类别中的分组步骤，建立排序记录步骤到分组步骤之间的连接，要求构成分组的字段为：客户 ID 和客户 TEL，聚合要求为订单数量是个数类型，支付总价为求和类型。

5. 根据客户订单消费金额进行排序，选择转换类别中的排序记录步骤，建立分组步骤到排序记录步骤之间的连接，设置按照“支付总价”字段降序排序。

任务三：数据的导出和维护（15分）

1. 选择输出类别中的 Microsoft Excel 输出步骤，建立排序记录 2 步骤到 Microsoft Excel 输出步骤之间的连接。（7分）
2. 配置 Microsoft Excel 输出步骤中相关参数，将清洗过的数据导出名为 03_1 outputdata.xlsx 的数据。（8分）

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-5 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	kettle 软件	用以开发程序
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经		测评专家满足任一条件

	验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。	
	结果测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。	

(4) 考核时量

考核时间为90分钟

(5) 评分标准

数据采集模块的考核实行100分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的20%，工作任务完成质量占该项目总分的80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	数据导入	数据导入正确	15分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	数据导入	数据导入正确	50分	
	数据的导出与维护	数据的导出与维护正确	15分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分		

24. 试题编号：3-6

(1) 任务描述

无人售货机在我们日常生活中已经较为的普遍，它是一种根据投币或者扫码支付的自动付货的机器，某无人售货机公司采集无人售货机订单信息数据（03_1_inputdata.csv）的部分关键字段说明如下表 1 所示，源数据文件类型为分隔符的数据类型，为了了解客户订单的状态，要求利用 Kettle 软件对该数据进行相应聚合计算、排序等操作，预览得到的数据结果如下图 1 所示。为了了解售货机每天销售的情况，要求利用 Kettle 软件对该数据进行过滤、字段选择、排序等清洗操作得到每台售货机每天的商品销售金额，预览得到的数据结果如下图 2 所示。

表 1 无人售货机订单信息数据字段说明

字段名称	位置	说明	示例
createdtime	1	订单生成的时间	2018/11/21 18:46:13
customerid	2	客户 ID	204524
customermobile	3	客户手机号码	18660951385
totalprice	4	订单总额	4.5
paytotalprice	5	订单实际支付金额	4.5
discounttotalprice	6	订单优惠金额	0
status	7	订单状态	SUCCESS
source	8	订单来源	ALIPAY_SHOPPING
boxid	9	售货机 ID	73216297342
ordernum	10	订单号	272074322789605000
payexceptiontype	11	订单异常菜单列表	COMMON
payedtime	12	支付时间	2018/11/21 18:46:14
productname	13	商品名称	康师傅经典红烧牛肉面
amount	14	商品数量	1
costprice	15	商品成本价	3.58
saleprice	16	商品销售价	4.5
productpaytotalprice	17	商品实际支付总金额	4.5

producttotalprice	19	商品支付总金额	4.5
-------------------	----	---------	-----

源数据文档名称：03_1_inputdata.csv



任务一：数据的导入（15分）

1. 在 Kettle 软件中新建名为 04-1 的转换文件（.ktr），选择输入类别中的 CSV 文件输入步骤。（7分）

2. 配置 CSV 文件输入步骤中的相关参数，将 03_1_inputdata.csv 中的源数据全部导入 kettle 平台中（注：<1>customermobile 列需改类型为 String；<2>createdtime 列需改类型为 String）。（8分）

任务二：数据的清洗（50分）

1. 对获取的数据进行过滤。选择流程类别中的过滤记录步骤，建立 CSV 文件输入步骤到过滤记录步骤之间的连接，设置参数，过滤掉商品名称为空和支付不成功的订单数据。（10分）

2. 对获取的数据进行抽取。选择转换类别中的字段选择步骤，建立过滤记录步骤到字段选择步骤之间的连接，只保留需要的 boxid、createdtime、amount 和 productpaytotalprice 字段，并且分别改为售货机 ID、订单生成时间、购买商品数量、商品实际支付总金额。（8分）

3. 对订单生成时间清洗为销售日期。选择转换类别中的剪切字符串步骤，建立字段选择步骤到剪切字符串步骤之间的连接，输入流字段为“订单生成时间”，输出流字段为“销售日期”，从 0 剪切到 10。（8分）

4. 根据售货机 ID 和销售日期进行排序，选择转换类别中的排序记录步骤，建立分剪切字符串骤到排序记录步骤之间的连接，设置按照“售货机 ID”和“销售日期”字段升序排序。（8分）

5. 对售货机的商品实际支付金额等字段进行分组聚合，统计各个售货机的日销售金额。选择统计类别中的分组步骤，建立排序记录步骤到分组步骤之间的连接，要求构成分组的字段为：售货机 ID 和销售日期，聚合要求为商品实际支付金额为求和类型。（8分）

6. 根据售货机 ID 邓进行排序，选择转换类别中的排序记录步骤，建立分组步骤到排序记录步骤之间的连接，设置按照“售货机 ID”和“销售日期”字段升序排序。（8分）

任务三：数据的导出和维护（15分）

1. 选择输出类别中的 Microsoft Excel 输出步骤，建立排序记录 2 步骤到 Microsoft Excel 输出步骤之间的连接。（7分）

2. 配置 Microsoft Excel 输出步骤中相关参数，将清洗过的数据导出名为 03_2outputdata.xlsx 的数据。（8分）

（2）提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-6 答案.docx”文件，文件内容格式为

“任务号+题号+命令详情+答案”，示例如下：

任务一（1）命令和截图：XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	kettle 软件	用以开发程序
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(4) 考核时量

考核时间为 90 分钟

(5) 评分标准

数据采集模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	数据导入	数据导入正确	15 分	1、考试舞弊、抄袭、没有按要求

	数据导入	数据导入正确	50 分	填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	数据的导出与维护	数据的导出与维护正确	15 分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分		

25. 试题编号：3-7

(1) 任务描述

某无人售货机公司采集无人售货机客户订单信息数据（03_1_inputdata.csv）和无人售货机信息数据（03_3_inputdata.csv）的部分关键字段说明如下表 1 和表 2 所示，客户订单详情数据记录着不同客户每天的每一笔订单详细数据，无人售货机信息数据记录着售货机的基本信息，从盈利的角度出发，为了了解售货机每天销售的情况，统计每台机器的利润，要求利用 Kettle 软件对该数据进行过滤、公式计算、排序等清洗操作。

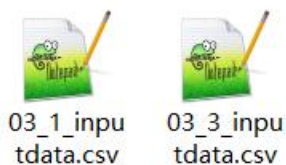
表 1 无人售货机订单信息数据字段说明

字段名称	位置	说明	示例
createdtime	1	订单生成的时间	2018/11/21 18:46:13
customerid	2	客户 ID	204524
customermobile	3	客户手机号码	18660951385
totalprice	4	订单总额	4.5
paytotalprice	5	订单实际支付金额	4.5
discounttotalprice	6	订单优惠金额	0
status	7	订单状态	SUCCESS
source	8	订单来源	ALIPAY_SHOPPING
boxid	9	售货机 ID	73216297342
ordernum	10	订单号	272074322789605000
payexceptiontype	11	订单异常菜单列表	COMMON
payedtime	12	支付时间	2018/11/21 18:46:14
productname	13	商品名称	康师傅经典红烧牛肉面
amount	14	商品数量	1
costprice	15	商品成本价	3.58
saleprice	16	商品销售价	4.5
productpaytotalprice	17	商品实际支付总金额	4.5
producttotalprice	19	商品支付总金额	4.5

表 2 无人售货机信息数据字段说明

字段名称	位置	说明	示例
boxid	1	售货机 ID	73216297360
address	2	售货机投放地址	山东省临沂市兰山区红旗路与通达路
name	3	售货机名称	城市风景 2#2 单元自产
qrcode	4	售货机二维码	http://assets.mayihezhi.com/prod/box/Qrcode/20189/21193050656de5cfc48-6ba3-4bac-8997-42455437bde2.jpg
serialnumber	5	售货机编码	A2448
status	6	售货机状态	ONLINE
modelnumber	7	售货机型号	MA650-A

源数据文档名称：03_1_inputdata.csv、03_3_inputdata.csv



任务一：数据的导入和转换（总共 20 分）

1. 在 Kettle 软件中新建名为 05_1 的转换文件（.ktr），选择输入类别中的 CSV 文件输入步骤。（5 分）

2. 配置 CSV 文件输入步骤中的相关参数，将 03_1_inputdata.csv 中的源数据全部导入 kettle 平台中。（5 分）

3. 再次选择输入类别中的 CSV 文件输入步骤，配置 CSV 文件输入 2 步骤中的相关参数，将 03_3_inputdata.csv 中的源数据全部导入 kettle 平台中。（10 分）

任务二：数据的清洗，处理后的数据示例如表（50分）

1. 对获取的数据进行过滤。选择流程类别中的过滤记录步骤，建立 CSV 文件输入步骤到过滤记录步骤之间的连接，设置参数，过滤掉支付不成功的数据，保留支付成功的数。

2. 对获取的数据进行抽取。选择转换类别中的字段选择步骤，建立过滤记录步骤到字段选择步骤之间的连接，只保留需一部分字段并且进行改名，具体如下图所示：

boxid	售货机ID
ordernum	订单号
amount	购买商品数量
productpaytotalprice	商品实际支付总金额
costprice	商品成本价
saleprice	商品销售价
productdiscountprice	商品优惠金额
paytotalprice	订单实际支付金额

3. 将售货机信息和订单数据关联起来。选择连接类别中的记录关联（笛卡尔输出）步骤，建立字段选择步骤和 CSV 文件输入 2 步骤到记录关联步骤之间的连接条件设为“售货机 ID=boxid”。

4. 修改和选择关联之后的售货机和订单数据。选择转换类别中的字段选择步骤，步骤改名为字段选择(关联数据)，建立记录关联步骤到字段选择(关联数据)步骤之间的连接，只保留需一部分字段并且进行改名，具体如下图所示：

boxid	售货机ID
name	售货机名称
address	售货机地址
订单号	
购买商品数量	
商品实际支付总金额	
商品优惠金额	
商品成本价	
商品销售价	
订单实际支付金额	

5. 计算订单中商品的利润。选择脚本类别中的公式组件，建立字段选择(关联数据)步骤到公式步骤之间的连接，设置新字段为“商品利润”，公式为“([商品实际支付总金额]-([购买商品数量]*[商品成本价])-[商品优惠金额])”，值的类型为 Number, 长度为 15。

任务三：数据的导出和维护（每题 5 分，总共 10 分）

1. 选择输出类别中的 Microsoft Excel 输出步骤，建立分组聚合步骤到 Microsoft Excel 输出步骤之间的连接，

2. 配置 Microsoft Excel 输出步骤中相关参数，将清洗过的数据导出名为 03_3_outputdata 的 xlsx 的数据。

(2) 提交要求

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-7 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX, 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(3) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	kettle 软件	用以开发程序

测评专家	现场测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。	测评专家满足任一条件
	结果测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。	

(4) 考核时量

考核时间为90分钟

(5) 评分标准

数据采集模块的考核实行100分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的20%，工作任务完成质量占该项目总分的80%。具体评价标准见下面描述：

评价内容		评分标准		备注
工作任务	数据导入	数据导入正确	20分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	数据导入	数据导入正确	50分	
	数据的导出与维护	数据的导出与维护正确	10分	
职业素养	专业素养	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分		

项目 3：基于 hive 的数据处理与分析

26. 试题编号：3-8

(1) 任务描述

某公司针对员工上班情况、薪资情况、扣薪情况进行调研，以出台针对提高员工上班效率和上班积极性的政策。该调研收集了员工的薪资详细信息

employess.txt，部门数据如下所示：

```
Lilith,30,6000,50,Finance Dept  
Byron,36,5000,25,Personnel Dept  
Yvette,21,4500,15.5,  
Arlen,28,8000,20,Finance Dept  
Rupert,39,10000,66,R&D Dept  
Deborah,41,6500,0,R&D Dept  
Tim,22,6000,36.5,Sales Dept  
Olga,36,5600,10,Sales Dept  
Bruno,43,6700,0,Personal Dept  
Flora,27,4000,35,Sales Dept
```

该数据文件包括员工姓名、员工年龄、员工薪资、迟到扣款和员工所属部门。请根据此数据，对该公司员工的情况进行分析。

任务一：表的创建与数据的导入（每题 10 分，总共 20 分）

1. 上传 employess.txt 的 HDFS 目录/hive/data/emp 下
2. 创建外部表 emp_table，有属性：
 - (1) 姓名 name，字符串
 - (2) 年龄 age，整型
 - (3) 薪资 salary，浮点型
 - (4) 迟到扣款 deduction，浮点型
 - (5) 所属部门 dept，字符串

将 employess.txt 数据加载到外部表 emp_table 中（不使用加载数据文件的操作），并进行验证。

任务二：数据的分析（每题 7.5 分，总共 60 分）

1. 计算每位员工日平均薪资，以每月 20 天工作日计算，显示结果中，日均工资那一列的列名为 day_salary。
2. 计算每位员工的年薪，显示结果中，年薪工资那一列的列名为 year_salary。

3. 计算公司员工的平均年龄。
4. 计算每个部门的平均薪资, 不计算 dept 为空值的部门。
5. 计算每个部门员工的平均年龄。
6. 查询员工薪资大于等于 5000, 小于等于 8000 的员工信息。
7. 查询销售部门 Sales Dept 中, 月薪大于等于 6000 的员工姓名、年龄
8. 查询员工薪资大于等于 5000, 小于等于 8000 的员工人数。

提交要求:

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹, 考生文件夹的命名规则: 考生学校+考生号+考生姓名, 示例: 永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-8 答案.docx”文件, 文件内容格式为“任务号+题号+命令详情+答案”, 示例如下:

任务一 (1) 命令和截图: XXXXXX, 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(2) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计, 每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家: 在本行业具有 3 年以上的从业经验 (工程师及以上职称) 或从事本专业具有 5 年以上的教学经验 (副高及以上职称), 或具有软件设计师、系统分析师、数据库设计师资格证书 (2 人/场)。		测评专家满足任一条件
	结果测评专家: 在本行业具有 3 年以上的从业经验 (工程师及以上职称) 或从事本专业具有 5 年以上的教学经		

验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。

（3）考核时量

90 分钟

（4）评分细则

数据清洗模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

评价内容		配分	评分标准		备注
工作任务	数据表的创建和数据导入	20分	数据上传	10分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			表创建	10分	
	数据的分析	60分	数据分析正确	60分	
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	

27. 试题编号：3-9

(1) 任务描述

高考在即，某学校收集了学生的意向大学数据、各科成绩数据 student_exam.txt，进行处理分析，以得出学生的学习情况和预计录取情况，下图是部分数据：

```
Mandy,Peking University-Wuhan University-Nankai University,Chemistry:90-Physics:98- Biology:83,126-135-140  
Jerome,Tsinghua University-Fudan University-Nanjing University,History:89-Politics:92- Geography:87,130-116-128  
Delia,Nanjing University-Wuhan University-Nankai University,Chemistry:87-Physics:95- Biology:73,102-123-112  
Ben,Tianjin Universit-Peking University-Fudan University,Chemistry:92-Physics:88-Biology:79,98-142-106
```

该数据文件包括学生姓名、意向大学、文综（政治、历史、地理）/理综（生物、化学、物理）成绩和综合（语、数、外）成绩，请进行数据分析。

任务一：表的创建与数据的导入（每题 10 分，总共 20 分）

1. 上传 student_exam.txt 的 HDFS 目录/hive/data/student_exam 下。（10 分）
2. 创建外部表 student_exam_table，有属性：（10 分）
 - (1) 姓名 name，字符串
 - (2) 意向大学 university， ARRAY<STRING>
 - (3) 文理综成绩 hu_sci MAP<STRING, FLOAT>
 - (4) 总和成绩 comp STRUCT<chinese:FLOAT, maths:FLOAT, english:FLOAT>

将 student_exam.txt 数据加载到外部表 student_exam_table 中（不使用加载数据文件的操作），并进行验证。

任务二：数据的分析（每题 7.5 分，总共 60 分）

1. 查询学生的第一志向大学
2. 查询学生的物理和历史成绩
3. 查询学生的数学成绩
4. 拆分学生成绩表 student_exam_table 中意向大学数据
5. 拆分学生成绩表 student_exam_table 中文理综成绩
6. 统计意向大学填写了 Peking University 的学生
7. 统计第一志向大学为 Tsinghua University 的学生人数
8. 统计数学成绩大于 120 的学生人数

提交要求：

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-9 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(2) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

90 分钟

(4) 评分细则

数据清洗模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

评价内容		配分	评分标准		备注
工作任务	数据表的创建和数据导入	20 分	数据上传	10 分	1、考试舞弊、抄袭、
			表创建	10 分	

	数据的分析	60分	数据分析正确	60分	没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	

28. 试题编号：3-10

(1) 任务描述

某公司针对员工上班情况、薪资情况、扣薪情况进行调研，以出台针对提高员工上班效率和上班积极性的政策。该调研收集了员工的薪资详细信息 employess.txt，部门数据如下所示：

```
Lilith,30,6000,50,Finance Dept
Byron,36,5000,25,Personnel Dept
Yvette,21,4500,15.5,
Arlen,28,8000,20,Finance Dept
Rupert,39,10000,66,R&D Dept
Deborah,41,6500,0,R&D Dept
Tim,22,6000,36.5,Sales Dept
Olga,36,5600,10,Sales Dept
Bruno,43,6700,0,Personal Dept
Flora,27,4000,35,Sales Dept
```

该数据文件包括员工姓名、员工年龄、员工薪资、迟到扣款和员工所属部门。请根据此数据，对该公司员工的情况进行分析。

任务一：表的创建与数据的导入（每题 10 分，总共 20 分）

1. 上传 employess.txt 的 HDFS 目录/hive/data/emp 下
2. 创建外部表 emp_table，有属性：
 - (1) 姓名 name，字符串
 - (2) 年龄 age，整型
 - (3) 薪资 salary，浮点型
 - (4) 迟到扣款 deduction，浮点型
 - (5) 所属部门 dept，字符串

将 employess.txt 数据加载到外部表 emp_table 中（不使用加载数据文件的操作），并进行验证。

任务二：数据的分析（每题 7.5 分，总共 60 分）

1. 查询部门分类
2. 查询各部门包含的员工数
3. 查询平均薪资大于 7000 的部门
4. 查询每个部门的最高、最低工资，不包括部门为空。

5. 将每个部门的工资从高到低排序
6. 薪资排名前 5 的员工信息
7. 查询年龄大于 30 岁的员工数量
8. 查询薪资大于等于 8000 的员工人数

提交要求:

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-10 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图： XXXXXX，并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(2) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

90 分钟

(4) 评分细则

数据清洗模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

评价内容		配分	评分标准		备注
工作任务	数据表的创建和数据导入	20分	数据上传	10分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			表创建	10分	
	数据的分析	60分	数据分析正确	60分	
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	

29. 试题编号：3-11

(1) 任务描述

现有某公司员工信息表 staff.txt，部门数据如下图，公司现需要分析员工的职级、部门等分布情况，以进行员工调整，使公司效益更大化。请进行数据分析。

```
7369,SMITH,CLERK,7902,1980-12-17,800,null,20
7499,ALLEN,SALESMAN,7698,1981-02-20,1600,300,30
7521,WARD,SALESMAN,7698,1981-02-22,1250,500,30
7566,JONES,MANAGER,7839,1981-04-02,2975,null,20,
7654,MARTIN,SALESMAN,7698,1981-09-28,1250,1400,30
7698,BLAKE,MANAGER,7839,1981-05-01,2850,null,30
7782,CLARK,MANAGER,7839,1981-06-09,2450,null,10
7788,SCOTT,ANALYST,7566,1987-04-19,3000,null,20
7839,KING,PRESIDENT,null,1981-11-17,5000,null,10
7844,TURNER,SALESMAN,7698,1981-09-08,1500,0,30
```

数据字段为：工 id, 员工名字, 工作岗位, 部门经理, 受雇日期, 薪水, 奖金, 部门编号，对应的英文名称为 id, name, job, mgr, hiredate, sal, bonus, deptid

任务一：表的创建与数据的导入（每题 10 分，总共 20 分）

1. 上传 staff.txt 的 HDFS 目录/hive/data/staff 下
2. 创建外部表 staff_table，有属性：

- (1) id 整型
- (2) name 字符串
- (2) job 字符串
- (3) mgr 整型
- (4) hiredate 字符串
- (5) sal 整型
- (6) bonus 整型
- (7) deptid 整型

将 staff.txt 数据加载到外部表 staff_table 中（不使用加载数据文件的操作），并进行验证。

任务二：数据的分析（每题 7.5 分，总共 60 分）

1. 查询奖金为空的员工信息
2. 计算不同部门经理的下级员工人数

3. 计算所有员工入职至今的工作月份数
4. 计算每个部门的平均薪资
5. 计算所有部门的平均奖金
6. 将员工的入职日期格式化成 ‘yyyy/MM/dd’ 的形式
7. 根据员工薪资，判断员工薪资级别。薪资小于 2000 的为 low，薪资大于等于 2000 并小于 5000 为 middle，否则返回 high。
8. 查询员工薪资大于等于 2000，小于等于 5000 的员工人数。

提交要求:

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-11 答案.docx”文件，文件内容格式为“任务号+题号+命令详情+答案”，示例如下：

任务一 (1) 命令和截图: XXXXXX, 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(2) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件

	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	
--	--	--

(3) 考核时量

90 分钟

(4) 评分细则

数据清洗模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

评价内容		配分	评分标准		备注
工作任务	数据表的创建和数据导入	20分	数据上传	10分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			表创建	10分	
	数据的分析	60分	数据分析正确	60分	
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	

30. 试题编号：3-12

(1) 任务描述

某公司针对员工上班情况、薪资情况、扣薪情况进行调研，以出台针对提高员工上班效率和上班积极性的政策。该调研收集了员工的薪资详细信息

employess.txt，部门数据如下所示：

```
Lilith,30,6000,50,Finance Dept
Byron,36,5000,25,Personnel Dept
Yvette,21,4500,15.5,
Arlen,28,8000,20,Finance Dept
Rupert,39,10000,66,R&D Dept
Deborah,41,6500,0,R&D Dept
Tim,22,6000,36.5,Sales Dept
Olga,36,5600,10,Sales Dept
Bruno,43,6700,0,Personal Dept
Flora,27,4000,35,Sales Dept
```

该数据文件包括员工姓名、员工年龄、员工薪资、迟到扣款和员工所属部门。

请根据此数据，对该公司员工的情况进行分析。

任务一：表的创建与数据的导入（每题 10 分，总共 20 分）

1. 上传 employess.txt 的 HDFS 目录/hive/data/emp 下
2. 创建外部表 emp_table，有属性：
 - (1) 姓名 name，字符串
 - (2) 年龄 age，整型
 - (3) 薪资 salary，浮点型
 - (4) 迟到扣款 deduction，浮点型
 - (5) 所属部门 dept，字符串

将 employess.txt 数据加载到外部表 emp_table 中（不使用加载数据文件的操作），并进行验证。

任务二：数据的分析（每题 7.5 分，总共 60 分）

1. 将列 salary 的查询结果的数据类型转换成 INT 类型
2. 去除员工姓名中的空格
3. 根据员工年龄，判断员工属于中年还是青年，年龄 ≥ 40 为中年 middle age，否则返回青年 youth。

4. 计算每个部门的平均薪资, 不计算 dept 为空值的部门。
5. 根据员工薪资, 判断员工薪资级别。薪资小于 5000 的为 low, 薪资大于等于 5000 并小于 8000 为 middle, 否则返回 high。
6. 将每个部门的工资从高到低排序。
7. 查询销售部门 Sales Dept 中, 月薪大于等于 6000 的员工姓名、年龄
8. 查询每个部门的最高、最低工资, 不包括部门为空。

提交要求:

1) 在“E:\技能抽查提交资料\”文件夹内创建考生文件夹, 考生文件夹的命名规则: 考生学校+考生号+考生姓名, 示例: 永州职业技术学院 01 张三。

2) 在考生文件夹中创建“试题编号 3-12 答案.docx”文件, 文件内容格式为“任务号+题号+命令详情+答案”, 示例如下:

任务一 (1) 命令和截图: XXXXXX, 并给出运行结果截图

3) 最终将考生文件夹进行压缩后提交。

(2) 实施条件

表 2-1-1 项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Centos7 或更高版本		用于程序设计, 每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Hadoop 分布式系统	用以项目开发
测评专家	现场测评专家: 在本行业具有 3 年以上的从业经验 (工程师及以上职称) 或从事本专业具有 5 年以上的教学经验 (副高及以上职称), 或具有软件设计师、系统分析师、数据库设计师资格证书 (2 人/场)。		测评专家满足任一条件
	结果测评专家: 在本行业具有 3 年以上的从业经验 (工程师及以上职称) 或从事本专业具有 5 年以上的教学经验 (副高及以上职称), 或具有软件设计师、系统分析		

师、数据库设计师资格证书（2人/场）。	
---------------------	--

(3) 考核时量

90 分钟

(4) 评分细则

数据清洗模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

评价内容		配分	评分标准		备注
工作任务	数据表的创建和数据导入	20分	数据上传	10分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			表创建	10分	
	数据的分析	60分	数据分析正确	60分	
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	

项目 4 Spark 大数据处理与分析

31. 试题编号：3-13

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一 Ubuntu 与 Windows 的共享文件夹搭建

1、在 Ubuntu 与 Windows 系统内分别创建 2 个文件夹，将 Windows 的文件夹在 Ubuntu 中加载为共享文件夹，在 Ubuntu 中输入相应的指令使共享文件夹生效，使得 Windows 和 Ubuntu 中文件能互通。

2、在 Windows 系统中创建一个 word.txt 文档，输入三行话(共计 9 单词)—Hadoop is good—Spark is Better—Spark is Fast.将这个 txt 文档传入 Ubuntu 中。

3、将命令和执行结果提交到指定位置。

任务二 Spark Local 模式的启动

1、输入相应的指令启动 spark-local 模式，进入 spark-scala 的交互式编程环境。

2、将命令和执行结果提交到答题卷的指定位置。

任务三 RDD 的创建

1、从内存中读取数据创建 RDD—先创建一个数组，数组元素为 1, 2, 3, 4. 输入相应的指令将该数组转为 RDD

2、从 Ubuntu 本地文件夹中读取相应数据创建 RDD, 输入相应指令读取任务一中的 word.txt 文档，将该文档转为 RDD

3、将命令和执行结果提交到指定位置。

(2) 实施条件

表 1-硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表 2-软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以	安装增强功能

		上	
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	伪分布式
6	Scala	2.12.15 及以上	不超过 2.13
7	Spark	3.2.0 及以上	Local 模式

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制,评价内容包括职业素养、工作任务完成情况两个方面。其中,职业素养占该项目总分的 20%,工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	Ubuntu与Windows的共享文件夹搭建	40分	成功在Ubuntu与Windows系统创建两个文件夹	5分	1、考试舞弊、抄袭、没有按要求填写相关信息,本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			正确加载Windows下为共享文件夹的操作和指令	15分	
			成功完成word.txt文档的创建	10分	
			成功完成word.txt文档成功传入Ubuntu中	10分	
	Spark Local模式的启动	10分	正确访问的spark安装路径,输入相应的指令启动spark-local模式,进入spark-scala的交互式编程环境	10分	
	RDD的创建	30分	正确的输入相应指令从内存中读取数据创建RDD	15分	
正确的输入相应指令从Ubuntu本地文件夹中读取相应数据生成RDD			15分		
职业素养	专业素养	10分	熟练使用相关软件,步骤命名规范,能做到见名知意,需要一定的注释进行解释说明。	0-10分	

养	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

32. 试题编号：3-14

(1) 任务描述

Spark RDD 提供了丰富的操作方法用于操作分布式的数据集合，包括转换操作和行动操作两部分。

转换操作可以将一个 RDD 转换为一个新的 RDD，但是转换操作是懒操作，不会立刻执行计算。行动操作是用于触发转换操作的操作，这时才会真正开始进行计算。而从内存中生成 RDD 一般用的 `parallelize()` 函数。

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一 使用 `map()` 方法转换数据

- 1、创建一个数组形式 RDD，数据形式-Array(1, 2, 3, 4)。
- 2、使用 `map()` 方法将这四元素各自加十。
- 3、将命令和执行结果提交到指定位置。

任务二 使用 `sortBy()` 方法进行排序

- 1、创建一个数组形式 RDD，数据形式-Array((1, 5), (45, 3), (7, 6))。
- 2、使用 `sortBy()` 方法根据元组的第一个值升序排序
- 3、使用 `sortBy()` 方法根据元组的第二个值降序排序。
- 4、将命令和执行结果提交到指定位置。

任务三 使用 `collect()`, `take()` 方法查询数据

- 1、使用 `collect()` 方法查询任务一和任务二中共计三个 RDD 的所有数据。
- 2、使用使用 `take()` 方法查询任务一和任务二中共计三个 RDD 的第一条数据
- 3、将命令和执行结果提交到指定位置。

任务四 使用 `union()` 方法合并多个 RDD

- 1、通过读取题目所提供的 `word1.txt`, `word2.txt` 两个 txt 文档创建 RDD。
- 2、使用 `union()` 方法合并这两个 RDD
- 3、使用 `collect()` 方法输出这两个 RDD 的数据。
- 4、将命令和执行结果提交到指定位置。

任务五 使用 `filter()` 方法进行过滤

- 1、使用 `filter()` 方法对任务四合并之后的 RDD 进行过滤。
- 2、通过 `contains()` 方法得到只包含 Spark 的数据
- 3、通过 `contains()` 方法得到只包含 Hadoop 的数据。

4、将命令和执行结果提交到指定位置。

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	
6	Scala	2.12.15 及以上	不超过 2.13
7	Spark	3.2.0 及以上	

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制, 评价内容包括职业素养、工作任务完成情况两个方面。其中, 职业素养占该项目总分的 20%, 工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	使用map()方法转换数据	10分	正确创建数组形式的RDD	5分	1、考试舞弊、抄袭、没有按要求填写相关信息, 本项目记0分。 2、严重违反考场纪律、造成恶劣影响
			正确使用map()方法将这四个元素各自加十	5分	
	使用 sortBy()方法 进行排序	20分	正确使用sortBy()方法根据元组的第一个值升序排序	10分	
			正确使用sortBy()方法根据元组的第二个值降序排序	10分	
	使用 collect(), take()	10分	正确使用collect()方法查询任务一和任务二中共计	5分	

	方法查询数据		三个RDD的所有数据	5分	的本项目记0分。
			正确使用使用take()方法查询任务一和任务二中共计三个RDD的第一条数据		
	使用 union() 方法合并多个 RDD	20分	成功读取 word1.txt, word2.txt 两个 txt 文档创建 RDD	10分	
			正确合并这两个 RDD 后用 collect() 查询数据	10分	
	使用 filter() 方法进行过滤	20分	正确使用相应函数过滤出包含 Hadoop 的数据	10分	
			正确使用相应函数过滤出包含 Spark 的数据	10分	
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

33. 试题编号：3-15

(1) 任务描述

Spark RDD 提供了丰富的操作方法用于操作分布式的数据集合，包括转换操作和行动操作两部分。

转换操作可以将一个 RDD 转换为一个新的 RDD，但是转换操作是懒操作，不会立刻执行计算。行动操作是用于触发转换操作的操作，这时才会真正开始进行计算。而从内存中生成 RDD 一般用的 `parallelize()` 函数。

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一 使用 `distinct()` 方法进行去重

- 1、创建一个数组形式 RDD，数据形式-`Array(('a', 1), ('b', 2), ('c', 3), ('a', 1), ('b', 1), ('a', 6), ('c', 1), ('b', 7), ('c', 5))`。
- 2、使用 `distinct()` 方法将 RDD 中重复元素去除。
- 3、将命令和执行结果提交到指定位置。

任务二 使用键值对 RDD 的 `reduceByKey()` 方法

- 1、创建一个数组形式 RDD，数据形式-`Array(('a', 1), ('b', 2), ('c', 3), ('a', 1), ('b', 1), ('a', 6), ('c', 1), ('b', 7), ('c', 5))`。
- 2、使用 `reduceByKey()` 方法和 `mapValues` 方法() 方法得到 a,b,c 三个 Key 值对应的 Value 值总和。
- 3、使用 `reduceByKey()` 方法和 `mapValues` 方法() 方法得到 a,b,c 三个 Key 值对应的 Value 值平均值。
- 4、将命令和执行结果提交到指定位置。

任务三 使用键值对 RDD 的 `groupByKey()` 方法

- 1、创建一个数组形式 RDD，数据形式-`Array(('a', 1), ('b', 2), ('c', 3), ('a', 1), ('b', 1), ('a', 6), ('c', 1), ('b', 7), ('c', 5))`。
- 2、使用 `groupByKey()` 方法对该 RDD 分组
- 3、使用 `mapValues` 方法() 计算 a,b,c 三个 Key 值对应的 Value 值总和和平均值。
- 4、将命令和执行结果提交到指定位置。

任务四 使用 `join()` 方法连接两个 RDD

- 1、创建两个数组形式 RDD，数据形式-`Array(('a', 1), ('b', 2), ('c', 3))`,

Array((' a', 1), (' d', 4), (' e', 5))。

- 2、使用 join() 方法连接这两个 RDD
- 3、使用 collect() 方法输出连接后 RDD 的数据。
- 4、将命令和执行结果提交到指定位置。

任务五 使用 zip() 方法组合将两个 RDD 组合成键值对 RDD

- 1、创建两个数组形式 RDD，数据形式-Array((' a', 1), (' b', 2), (' c', 3)), Array((' a', 1), (' d', 4), (' e', 5))。
- 2、使用 zip() 方法将这两个 RDD 合并为键值对形式
- 3、将命令和执行结果提交到指定位置。

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	
6	Scala	2.12.15 及以上	不超过 2.13
7	Spark	3.2.0 及以上	

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述：

Spark 大数据处理与分析模块考核评价标准

评价内容	配	评分标准	备注
------	---	------	----

		分			
工作任务	使用distinct()方法进行去重	10分	正确创建一个数组形式RDD	5分	1、考试舞弊、抄袭、没有按要求填写相关信息,本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			正确使用distinct()方法将RDD中重复元素去除	5分	
	使用键值对RDD的reduceByKey()方法	20分	正确使用reduceByKey()方法得到对应的Value值总和	10分	
			正确使用reduceByKey()方法得到对应的Value值平均值	10分	
	使用键值对RDD的groupByKey()方法	20分	正确使用groupByKey()方法对该RDD分组	10分	
			正确使用mapValues方法()计算对应的Value值总和和平均值。	10分	
	使用join()方法连接两个RDD	15分	成功创建RDD	5分	
			正确使用join()方法连接这两个RDD	10分	
	使用zip()方法组合将两个RDD组合成键值对RDD	15分	创建RDD	5分	
			正确使用zip()方法将这两个RDD合并为键值对形式	10分	
职业素养	10分	熟练使用相关软件,步骤命名规范,能做到见名知意,需要一定的注释进行解释说明。		0-10分	
		着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。		0-10分	
总计		100分			

34. 试题编号: 3-16

(1) 任务描述

天气预报每天都会显示各城市得温度，方便出行人士根据当天的温度穿上合适的衣服。现有一份各城市的温度数据文件 avgTemperature.txt，数据如下表所示，记录了某段时间范围内各城市每天的温度，文件中每一行数据分别表示城市名和温度，现要求使用 Spark 编程计算出各城市的平均温度。

表 3-8 各城市温度部分数据

beijing	28.1
shanghai	28.7
guangzhou	32.0
shenzhen	33.1
beijing	27.3
shanghai	30.1
guangzhou	33.3
shenzhen	28.6
beijing	28.2
shanghai	29.1
guangzhou	32.0
shenzhen	32.1

以下所有任务的答案、截图、文件等，保存到物理机上指定位置----“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一 数据文件上传到 HDFS 上，读取 HDFS 上的数据文件并创建 RDD

- 1、将放置于 Windows 系统下的数据文件通过共享文件夹传到 Ubuntu 系统下
- 2、启动 Hadoop
- 3、使用相应指令或图形化操作界面在 HDFS 创建该温度数据文件的存放位置文件夹
- 4、使用相应指令或图形化操作界面将位于 Ubuntu 内部温度数据文档上传到 HDFS。
- 5、输入相应指令使用 textFile () 方法读取 HDFS 上的温度数据创建 RDD
- 6、将命令和执行结果提交到指定位置。

任务二 对原始数据 RDD 进行处理得到各个城市的平均温度

- 1、将原始数据按制表符进行转化为(城市，温度)的键值对形式。
- 2、输入相应指令给温度数据按城市分组，得到每个城市对应的所有温度。
- 3、求出每个城市对应的所有温度数据的平均值。
- 4、将命令和执行结果提交到指定位置。

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	伪分布式
6	Scala	2.12.15 及以上	不超过 2.13
7	Spark	3.2.0 及以上	Local 模式

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制, 评价内容包括职业素养、工作任务完成情况两个方面。其中, 职业素养占该项目总分的 20%, 工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	数据文件上传到HDFS上, 读取HDFS上的数据文件并创建RDD	20分	成功将Windows系统下的数据文件传到Ubuntu系统	5分	1、考试舞弊、抄袭、没有按要求填写相关信息, 本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			成功将位于Ubuntu内部温度数据文档上传到HDFS。	5分	
			正确读取HDFS上温度数据创建RDD	10分	
	对原始数据RDD进行处理得到各个城市的平均温度	60分	成功将原始数据RDD 转化为(城市, 温度)形式	20分	
			正确将温度数据按城市分组, 得到每个城市对应的所有温度	20分	

			正确求出每个城市对应的所有温度数据的平均值	20分	
职业素养	专业素养	10分	熟练使用相关软件,步骤命名规范,能做到见名知意,需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10分	
总计		100分			

35. 试题编号：3-17

(1) 任务描述

2021年,某公司为了提高员工工作的积极性,将对公司员工进行一次调薪,

需要根据员工在 2020 年的薪资情况及在职表现重新调整薪资，对于爱岗敬业的公司员工，公司拟根据其业绩分析情况予以不同程度涨薪。员工 2020 年上半年薪资文件（Employee_salary_first_half.csv）表头信息如下所示。

要求：员工 2020 年上半年薪资文件，利用 Spark 技术查询上半年实际薪资排名前 3 的员工信息。

表 1 字段信息表

字段名称	说明	字段名称	说明
EmpID	员工ID	GROSS	总薪资
Name	姓名	Net_Pay	实际薪资
Gender	性别	Deduction	薪资扣除部分
Date_of_Birth	出生日期	Designation	职位
Age	年龄	Department	部门

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：读取员工薪资数据创建 RDD。

- 1、将放置于 Windows 系统下的数据文件通过共享文件夹传到 Ubuntu 系统下
- 2、启动 Hadoop，启动 Spark
- 3、使用相应指令或图形化操作界面在 HDFS 创建该薪资数据文件的存放位置文件夹
- 4、使用相应指令或图形化操作界面将位于 Ubuntu 内部温度数据文档上传到 HDFS。
- 5、输入相应指令使用 `textFile()` 方法读取数据创建 RDD
- 6、将命令和执行结果提交到指定位置。

任务二：去除原始数据的表头信息，并获取第二列员工姓名和第七列薪水数据。

- 1、将表头信息去除。
- 2、将数据按分隔符“,”分隔。
- 3、取出第 2 列员工姓名和第 7 列实际薪资数据，并将实际薪资数据转换成 `Int` 类型数据。

任务三：获取上半年实际薪资排名前 3 的员工信息。

- 1、根据实际薪资进行降序排列。
- 2、获取上半年实际薪资排名前3的员工信息。

(2) 实施条件

表2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	伪分布式
6	Scala	2.12.15 及以上	不超过 2.13
7	Spark	3.2.0 及以上	Local 模式

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制, 评价内容包括职业素养、工作任务完成情况两个方面。其中, 职业素养占该项目总分的 20%, 工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	数据文件上传到HDFS上, 读取HDFS上的数据文件并创建RDD	20分	Windows系统下的数据文件传到Ubuntu系统	5分	1、考试舞弊、抄袭、没有按要求填写相关信息, 本项目记0分。 2、严重违反考场
			位于Ubuntu内部薪资数据数据文档上传到HDFS。	5分	
			读取数据创建RDD	10分	

	获取第二列员工姓名和第七列薪水数据	40分	表头信息去除	20分	纪律、造成恶劣影响的本项目记0分。
			数据按分隔符“,”分隔	20分	
			将数据转为(姓名, 薪资)形式, 薪资为Int类型		
	获取上半年实际薪资排名前3的员工信息	20分	薪水降序排序	10分	
取出前三员工信息			10分		
职业素养	专业素养	10分	熟练使用相关软件, 步骤命名规范, 能做到见名知意, 需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明, 遵守考场纪律, 按顺序进出考场。	0-10分	
总计		100分			

36. 试题编号：3-18

(1) 任务描述

2021年, 某公司为了提高员工工作的积极性, 将对公司员工进行一次调薪, 需要根据员工在2020年的薪资情况及在职表现重新调整薪资, 对于爱岗敬业的公司员工, 公司拟根据其业绩分析情况予以不同程度涨薪。员工2020年下半年薪资文件(Employee_salary_second_half.csv)表头信息如下所示。

要求: 员工2020年下半年薪资文件, 利用Spark技术查询下半年实际薪资最低的10位员工姓名。

表 1 字段信息表

字段名称	说明	字段名称	说明
EmpID	员工ID	GROSS	总薪资
Name	姓名	Net_Pay	实际薪资
Gender	性别	Deduction	薪资扣除部分
Date_of_Birth	出生日期	Designation	职位
Age	年龄	Department	部门

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：读取员工薪资数据创建 RDD。

- 1、将放置于 Windows 系统下的数据文件通过共享文件夹传到 Ubuntu 系统下
- 2、启动 Hadoop，启动 Spark
- 3、使用相应指令或图形化操作界面在 HDFS 创建该薪资数据文件的存放位置文件夹
- 4、使用相应指令或图形化操作界面将位于 Ubuntu 内部温度数据文档上传到 HDFS。
- 5、输入相应指令使用 `textFile()` 方法读取数据创建 RDD
- 6、将命令和执行结果提交到指定位置。

任务二：去除原始数据的表头信息，并获取第二列员工姓名和第七列薪水数据。

- 1、将表头信息去除。
- 2、将数据按分隔符“,”分隔。
- 3、取出第 2 列员工姓名和第 7 列实际薪资数据，并将实际薪资数据转换成 `Int` 类型数据。

任务三：获取下半年实际薪资最低的 10 位员工姓名。

- 1、根据实际薪资进行升序排列。
- 2、获取上半年实际薪资最低的 10 位员工姓名。

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注

1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网
---	-----	-----	--------------------------------------	-------

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	伪分布式
6	Scala	2.12.15 及以上	不超过 2.13
7	Spark	3.2.0 及以上	Local 模式

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制, 评价内容包括职业素养、工作任务完成情况两个方面。其中, 职业素养占该项目总分的 20%, 工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	数据文件上传到HDFS上, 读取HDFS上的数据文件并创建RDD	20分	成功将Windows系统下的数据文件传到Ubuntu系统	5分	1、考试舞弊、抄袭、没有按要求填写相关信息, 本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			成功将位于Ubuntu内部薪资数据数据文档上传到HDFS。	5分	
			成功将读取数据创建RDD	10分	
	获取第二列员工姓名和第七列薪水数据	40分	成功去除表头信息	20分	
			数据按分隔符“,”分隔 正确将数据转为(姓名, 薪资)形式, 薪资为Int类型	20分	

	获取上半年实际薪资排名前3的员工信息	20分	正确升序排序薪水	10分	
			正确取出倒数前十的员工姓名	10分	
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

37. 试题编号：3-19

(1) 任务描述

2021年，某公司为了提高员工工作的积极性，将对公司员工进行一次调薪，需要根据员工在2020年的薪资情况及在职表现重新调整薪资，对于爱岗敬业的公司员工，公司拟根据其业绩分析情况予以不同程度涨薪。公司有员工2020年上半年薪资文件（Employee_salary_first_half.csv）和下半年薪资文件（Employee_salary_second_half.csv），两份文件的数据格式和数据字段均相同，以员工2020年上半年的薪资文件 Employee_salary_first_half.csv 为例，文件共有10个数据字段。

要求：根据员工2020年上、下半年两份薪资文件，利用Spark技术统计每一位员工2020年的全年薪资情况，月均薪资情况，并将员工上下半年薪资情况，

全年薪资情况，月均薪资情况。

表 1 字段信息表

字段名称	说明	字段名称	说明
EmpID	员工ID	GROSS	总薪资
Name	姓名	Net_Pay	实际薪资
Gender	性别	Deduction	薪资扣除部分
Date_of_Birth	出生日期	Designation	职位
Age	年龄	Department	部门

以下所有任务的答案、截图、文件等，保存到物理机上指定位置----“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：读取员工上下半年的两份薪资数据并进行处理。

- 1、将位于 Windows 下两份薪资数据传至 HDFS。
- 2、构建读取两份薪资数据后去表头按然后分隔后的 RDD。
- 3、构建两份薪资数据的第 2 列员工姓名和第 7 列实际薪资数据 RDD，并将实际薪资数据转换成 Int 类型数据。
- 4、将命令和执行结果提交到指定位置。

任务二：合并员工上下半年的两份薪资数据，处理数据得到员工全年薪水。

- 1、合并员工上下半年的两份薪资数据。
- 2、将员工上下半年的薪资相加，得到员工全年薪资的 RDD。
- 3、把员工全年薪资数据降序排列，输出数据。
- 4、将命令和执行结果提交到指定位置。

任务三：处理员工全年薪资数据，得到员工月均薪资。

- 1、合并员工上下半年的两份薪资数据。
- 2、将员工上下半年的薪资相加，得到员工全年薪资的 RDD。
- 3、使用 `groupByKey()` 方法对员工全年薪资 RDD 分组。
- 4、使用 `mapValues()` 方法对键值对 RDD 的 Values 进行计算得到月工资。
- 5、将命令和执行结果提交到指定位置。

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	伪分布式
6	Scala	2.12.15 及以上	不超过 2.1.3
7	Spark	3.2.0 及以上	Local 模式

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制, 评价内容包括职业素养、工作任务完成情况两个方面。其中, 职业素养占该项目总分的 20%, 工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	读取员工上下半年的两份薪资数据并进行处理	20 分	成功将位于Windows下两份薪资数据传至HDFS	5 分	1、考试舞弊、抄袭、没有按要求填写相关信息, 本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项
			成功薪资数据后去表头按然后分隔RDD	5 分	
			成功构建(员工姓名, 薪资)键值对形式的RDD	10 分	
	合并员工上下半年的两份薪	30 分	成功上下半年员工薪资合并	10 分	

	资数据, 处理数据得到员工全年薪水		成功计算全年员工薪资	10分	目记0分。
			成功降序排序员工整年薪资	10分	
	处理员工全年薪资数据, 得到员工月均薪资	30分	groupBykey() 方法对员工全年薪资RDD分组	15分	
			使用mapValues() 方法进行计算得到月工资	15分	
职业素养	专业素养	10分	熟练使用相关软件, 步骤命名规范, 能做到见名知意, 需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明, 遵守考场纪律, 按顺序进出考场。	0-10分	
总计		100分			

38. 试题编号：3-20

(1) 任务描述

Spark 编程的基础操作的过程是在 spark-shell 的交互式环境中进行的, 这样的交互式环境会对每个指令做出反馈, 适合初学者学习或调试代码时使用。

而在真实的生产环境中, 完成一个任务通常需要很多行代码, 并且需要多个类协作才能实现, 因此需要更加适合的开发环境, 如 IntelliJ IDEA。

要求: 在 Ubuntu 中安装 IntelliJ IDEA。在 IntelliJ IDEA 中安装 Scala 插件, 添加 Spark 开发依赖包, 配置 Spark 运行环境, 实现 Spark 工程的创建。

Spark 程序的编写与运行。

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：在 Ubuntu 中安装 Linux 版本 IDEA。

- 1、将 Linux 版本 IDEA 压缩包上传至 Ubuntu。
- 2、在 Ubuntu 中安装 IDEA，并添加 IDEA 程序到启动栏里快捷启动。
- 3、IDEA 中 Scala 插件安装。
- 4、将命令和执行结果提交到指定位置。

任务二：Maven 软件安装及配置。

1. 在 IDEA 中的 Project Structure 配置里的 Global Libraries 导入相应版本的 Scala(推荐 2.12.15)
2. 在 pom.xml 文档中补全 Maven 依赖
3. IDEA 中正确配置 Maven 工程, 编写“打印 hello, world”的 Scala 程序, 运行测试, 能运行则代表配置成功。
4. 将命令和执行结果提交到指定位置。

任务三：完成 WordCount 词频统计程序开发。

1. 将 words.txt 文档传入 Ubuntu 中。
2. 在 IDEA 中补全 Spark-jars 依赖。
3. IDEA 中正确配置 Maven 工程, 补全 Scala 程序代码, 完成词频统计。
4. 将命令和执行结果提交到指定位置。

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能

4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	
6	Scala	2.12.15 及以上	不超过 2.1.3
7	Spark	3.2.0 及以上	

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制,评价内容包括职业素养、工作任务完成情况两个方面。其中,职业素养占该项目总分的 20%,工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	在Ubuntu中安装Linux版本IDEA	15分	成功将Linux版本IDEA压缩包上传至Ubuntu	5分	1、考试舞弊、抄袭、没有按要求填写相关信息,本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
			成功在Ubuntu中安装IDEA	5分	
			成功在IDEA中安装Scala插件	5分	
	Maven软件安装及配置	20分	成功在本地Scala导入	5分	
			成功在pom.xml文档中补充Maven依赖	5分	
			Scala程序测试Maven配置是否成功。	10分	
完成WordCount词频统计程序开发	45分	成功将words.txt文档传入Ubuntu中	5分		
		成功在IDEA中补充Spark-jars依赖	10分		
		成功编写Scala程序,完成词频统计代码	30分		
职业素养	专业素养	10分	熟练使用相关软件,步骤命名规范,能做到见名知意,需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10分	

总计	100 分
----	-------

39. 试题编号：3-21

(1) 任务描述

Spark SQL 是 Spark 用于处理结构化数据的 Spark 模块，SparkSQL 不但兼容 Hive，还可以从 RDD、parquet 文件、JSON 文件中 获取数据，以后甚至支持获取 RDBMS 数据以及 cassandra 等 NOSQL 数据。

要求：在虚拟机中安装 MySQL 并配置密码。现有一份 employee.json 数据，利用 SparkSQL 技术对所提供数据进行探索与分析。

```
{"id":1,"name":"Ella","age":36}
{"id":2,"name":"Bob","age":29}
{"id":3,"name":"Jack","age":29}
{"id":4,"name":"Tom","age":29}
{"id":5,"name":"Petter","age":55}
```

图 1 数据部分展示

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：在 Ubuntu 中安装 MySQL 配置 root 用户密码为 123456。

- 1、Ubuntu 中安装 MySQL。
- 2、配置 root 用户密码为 123456。
- 3、启动 MySQL，在 MySQL 数据库中新建数据库 sparktest，再建表 employee，插入下列两行数据。

id	name	gender	age
1	Alice	F	22
2	John	M	25

任务二：Spark SQL 操作-DataFrame 的创建。

- 1、读取所提供的 employee.json 文件。
- 2、读取数据生成 DataFrame（以下方式三选一）
 - (1) 利用反射机制读取 RDD 生成 Dataframe，具体操作为：定义样例类 (Case Class)，根据例类 (Case Class) 修改相应数据类型并生成 DataFrame。
 - (2) 使用编程方式定义 RDD 模式生成 DataFrame，具体操作为：读取文件生成 RDD；定义模式信息 (Schema)；根据模式信息 (Schema) 完成相应信息生成 DataFrame。
 - (3) 使用读取文件的方式生成 DataFrame。
- 3、数据类型显示为 DataFrame 既完成该任务，任意方法都可。

任务三：Spark SQL 数据分析。

- 1、先查询所有数据，然后去除重复的数据。
- 2、查询年龄 age 的平均值，最小值和最大值。
- 3、数据按 name 分组，然后按 name 升序排列。

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存	要求能上网

			16GB 以上, 硬盘 320G	
--	--	--	------------------	--

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	伪分布式
6	Scala	2.12.15 及以上	不超过 2.1.3
7	Spark	3.2.0 及以上	Local 模式

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制, 评价内容包括职业素养、工作任务完成情况两个方面。其中, 职业素养占该项目总分的 20%, 工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	在Ubuntu中安装MySQL	20分	1、成功安装MySQL。 2、能以root用户登录。 3、成功创建employee表和插入数据	20分	1、考试舞弊、抄袭、没有按要求填写相关信息, 本项目记0分。 2、严重违反考场纪律、造成恶劣影响的, 本项目记0分。
	Spark SQL基本操作-DataFrame的创建	30分	正确将提供的employee.json文件读取完成后, 数据类型显示为DataFrame	30分	
	Spark SQL数据分析	30分	1、成功查询数据然后去重。 2、成功查询年龄age的平均值, 最小值和最大值 3、成功将数据按name分组, 然后按按name升序排列	30分	
职业素养	专业素养	10分	熟练使用相关软件, 步骤命名规范, 能做到见名知意, 需要一定的注释进行解释说	0-10分	

			明。		
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

40. 试题编号：3-22

(1) 任务描述

DataFrame 的推出，让 Spark 具备了处理大规模结构化数据的能力，不仅比原有的 RDD 转化方式更加简单易用，而且获得了更高的计算性能。现有学生考试成绩信息文件 `students.dat`、`students.csv`（两份文件数据内容一致，格式不同），表头信息为：班级-姓名-年龄-性别-科目-成绩。将这份数据读取为 DataFrame 的形式并分析。数据部分信息如下所示，

要求：根据提供的数据，将这份数据读取为 DataFrame 的形式，利用 Spark 技术对学生数据进行探索与分析。

```
12 张三 25 男 chinese 50
12 张三 25 男 math 60
12 张三 25 男 english 70
12 李四 20 男 chinese 50
12 李四 20 男 math 50
12 李四 20 男 english 50
12 王芳 19 女 chinese 70
12 王芳 19 女 math 70
12 王芳 19 女 english 70
13 张大三 25 男 chinese 60
13 张大三 25 男 math 60
13 张大三 25 男 english 70
13 李大四 20 男 chinese 50
13 李大四 20 男 math 60
13 李大四 20 男 english 50
13 王小芳 19 女 chinese 70
13 王小芳 19 女 math 80
13 王小芳 19 女 english 70
```

图 1 数据部分展示

以下所有任务的答案、截图、文件等，保存到物理机上指定位置——“考场说明指定路径\文件夹内创建考生文件夹\”。考生文件夹的命名规则：考生学校+Spark 大数据处理与分析+考生号+考生姓名，示例：永州职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：读取数据生成 DataFrame（以下方式三选一）

1、利用反射机制读取 RDD 生成 Dataframe，具体操作为：定义样类（Case Class），根据例类(Case Class)修改相应数据类型并生成 DataFrame。

2、使用编程方式定义 RDD 模式生成 DataFrame，具体操作为：读取文件生成 RDD；定义模式信息(Schema)；根据模式信息(Schema)完成相应信息生成 DataFrame。

3、使用读取文件的方式生成 DataFrame。

任务二：使用 DSL（领域专用语言）或 SQL（数据库）完成学生数据分析。

- 1、参加考试的总人数
- 2、年龄小于 20 岁的参加考试的人数
- 3、十二班参加考试的人数

任务三：使用 DSL（领域专用语言）或 SQL（数据库）完成学生成绩数据分析。

- 1、每个学生的平均成绩

2、十二班科目每科平均成绩

3、十三班科目每科平均成绩

(2) 实施条件

表 2 硬件环境

序号	设备	数量	规格	备注
1	计算机	1 台	CPU Intel 酷睿 i7, 内存 16GB 以上, 硬盘 320G	要求能上网

表 3 软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7 以上	安装 64 位版本
2	VirtualBox	6.1 或以上	安装在 64 位操作系统中
3	Ubuntu	22.04.2 LTS 版本及以上	安装增强功能
4	JDK	1.8.0	任意 JDK8 版本
5	Hadoop	3.1.3	伪分布式
6	Scala	2.12.15 及以上	不超过 2.1.3
7	Spark	3.2.0 及以上	Local 模式

(3) 考核时量

90 分钟。

(4) 评分标准

Spark 大数据处理与分析模块的考核实行 100 分制, 评价内容包括职业素养、工作任务完成情况两个方面。其中, 职业素养占该项目总分的 20%, 工作任务完成质量占该项目总分的 80%。具体评价标准见下面描述:

Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	数据文件读取为 DataFrame	20 分	文件数据正确读取为 DataFrame	20 分	1、考试舞弊、抄袭、没有按要求填写相关信息, 本项目记 0 分。
	使用 DSL (领域专用语言) 或 SQL (数据库) 完成学生数据分析	30 分	正确分析数据, 每题 10 分	30 分	

	使用DSL（领域专用语言）或SQL（数据库）完成学生成绩数据分析	30分	正确分析数据，每题10分	30分	2、严重违反考场纪律、造成恶劣影响的本项目记0分。
职业素养	专业素养	10分	熟练使用相关软件，步骤命名规范，能做到见名知意，需要一定的注释进行解释说明。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

模块四 数据可视化

项目 1: matplotlib 数据可视化

41. 试题编号：4-1

(1) 项目描述

2021 年的某天电影票房的具体数据保存在文件 data01.csv 中。现要求你根据所提供的数据文件，通过 **pandas** 工具读取数据文件，完成相关图表的绘制。

```

排序, 影片名称, 单日票房(万), 环比变化, 累计票房(万), 平均票价, 场均人次, 口碑指数, 上映天数
1, 当男人恋爱时, 929, -37%, 11269, 31, 6, 6, 96, 11
2, 了不起的老爸, 730, -66%, 6349, 36, 3, 7, 01, 4
3, 守岛人, 389, -44%, 3428, 40, 4, -, 4
4, 你好世界, 322, -41%, 8442, 30, 4, 7, 27, 11
5, 黑白魔女库伊拉, 276, -53%, 11274, 34, 4, 6, 91, 16
6, 超越, 256, -46%, 12705, 34, 2, 6, 16, 10
7, 阳光姐妹淘, 206, -34%, 8002, 35, 3, 4, 79, 11
8, 比得兔2逃跑计划, 180, -82%, 11220, 31, 2, 7, 39, 11
9, 寂静之地, 2149, -32%, 23777, 34, 4, 6, 87, 25
10, 困在时间里的父亲, 132, -66%, 917, 36, 6, 8, 27, 4

```

图 4-1-1 文件内容

(2) 任务描述

1. 导入数据分析和可视化需用到的相关模块，其中包括完成下列①和②中要求的导入操作。

①使用 `import` 语句导入 `matplotlib.pyplot` 并取别名为 `plt`。

②使用 `import` 语句导入 `pandas` 并取别名为 `pd`。

2. 通过 `pandas` 中的 `read_csv()` 函数读取数据文件中的内容，并通过设置参数 `encoding` 的值为 `'gbk'`，实现中文的正确读取。然后筛选出绘图所需的数据列。最后利用 `matplotlib` 库绘制横向柱状图。（图表的颜色可以采用默认值）。

3. 请将 `pyplot` 中的 `rc` 参数 `font.sans-serif` 的值设置为 `"SimHei"`，`axes.unicode_minus` 的值设成 `False`。

4. 横向柱状图的标题为“单日票房统计”。

5. 横向柱状图横坐标为“票房（万）”。

6. 横向柱状图纵坐标为影片名称，如图 4-1-2 所示。

7. 将绘图函数 `barh()` 中的参数 `height` 设置为 `0.5`。

8. 将所绘制的横向柱状图利用 `savefig()` 函数保存到与源代码相同的目录下，文件名为 `"fig01.png"`。

9. 使用 `show()` 函数显示上述绘制的图表。

10. 示例如下。

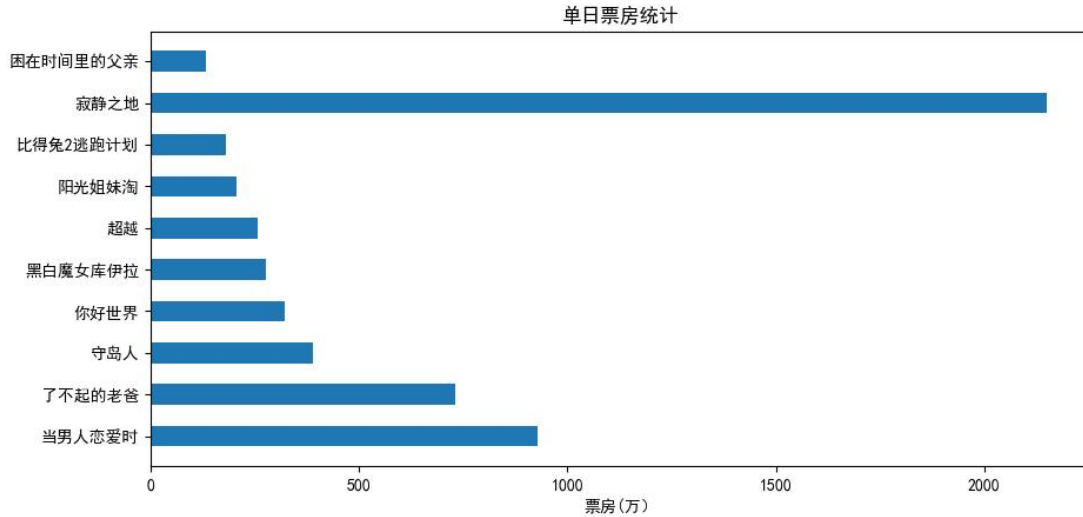


图 4-1-2 单日票房统计

提交要求:

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-1-1 数据可视化模块项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析		测评专家满足任一条件

	师、数据库设计师资格证书（2人/场）。	
	结果测评专家：在本行业具有3年以上的从业经验（工程师及以上职称）或从事本专业具有5年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2人/场）。	

（3）考核时量

考核时间为90分钟

（4）评分标准

数据分析与可视化模块的考核实行100分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的20%，工作任务完成质量占该项目总分的80%。具体评价标准见下表：

表 4-1-2 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10分	导入相关库是否正确	10分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影响的，本项目记0分。
	设置 rc 参数	5分	rc 参数设置是否正确	5分	
	读文件及筛选数据	10分	文件和筛选数据是否获取成功	10分	
	绘制图形	15分	绘制图形是否符合要求	15分	
	设置标题	10分	设置图表标题是否符合要求	10分	
	设置横坐标	10分	设置图表横坐标是否符合要求	10分	
	设置纵坐标	10分	设置图表纵坐标是否符合要求	10分	
	保存和显示图表	10分	保存和显示图表是否符合要求	10分	
职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释	0-10分	

			规范。		
	道德规范	10 分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100 分			

42. 试题编号：4-2

(1) 任务描述

2021 年的某周票房具体数据保存在文件 data02.csv 中。现要求你根据所提供的数据文件，通过 **pandas** 工具读取数据文件，完成相关图表的绘制。

排序, 影片名称, 单周票房 (万), 环比变化, 累计票房 (万), 平均票价, 场均人次, 口碑指数, 上映天数

- 1, 悬崖之上, 50047, 97%, 76125, 39, 18, 7. 79, 10
- 2, 你的婚礼, 31120, -27%, 73615, 38, 13, 5. 54, 10
- 3, 扫黑决战, 17589, 271%, 22358, 34, 14, 6. 51, 9
- 4, 追虎擒龙, 1526, 27%, 20570, 38, 10, 5. 68, 9
- 5, 秘密访客, 8188, -35%, 20837, 38, 7, 5. 87, 9
- 6, 猪猪侠大电影恐龙日记, 3767, 44%, 6381, 31, 9, -, 9
- 7, 名侦探柯南: 绯色的子弹, 1908, -21%, 21152, 35, 10, 6. 28, 23
- 8, 哥斯拉大战金刚, 1513, -14%, 122739, 36, 10, 7. 09, 45
- 9, 阳光劫匪, 1253, -58%, 4253, 36, 4, 5. 17, 9
- 10, 真三国无双, 590, -37%, 1524, 38, 4, -, 9

图 4-2-1 文件内容

(2) 任务要求

1. 导入数据分析和可视化需用到的相关模块, 其中包括完成下列①和②中要求的导入操作。

①使用 `import` 语句导入 `matplotlib.pyplot` 并取别名为 `plt`。

②使用 `import` 语句导入 `pandas` 并取别名为 `pd`。

2. 通过 `pandas` 中的 `read_csv()` 函数读取数据文件中的内容, 并通过设置参数 `encoding` 的值为 `'gbk'`, 实现中文的正确读取。然后筛选出绘图所需的数据列。

最后利用 `matplotlib` 库绘制柱状图。(图表的颜色可以采用默认值)。

3. 请将 `pyplot` 中的 `rc` 参数 `font.sans-serif` 的值设置为 `"SimHei"`, `axes.unicode_minus` 的值设成 `False`。

4. 柱状图的标题为 `"单周票房统计"`。

5. 柱状图纵坐标为 `"票房 (万)"`。

6. 柱状图横坐标为影片名称。

7. 将绘图函数 `bar()` 中的参数 `width` 设置为 `0.5`。

8. 将 `xticks()` 函数中的参数 `rotation` 设置为 `30`, 实现横坐标中的影片名称倾斜 `30°`。如图 4-2-2 所示。

9. 将所绘制的柱状图利用 `savefig()` 函数保存到与源代码相同的目录下, 文件名为 `"fig02.png"`。

10. 使用 `show()` 函数显示上述绘制的图表。

11. 示例如下。

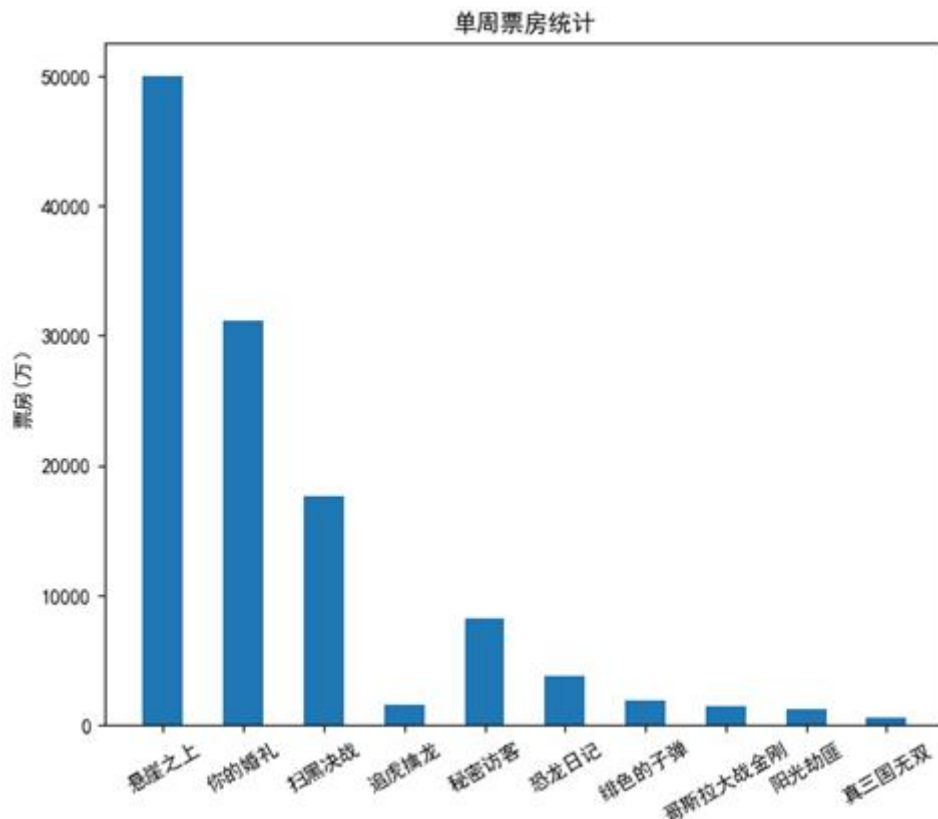


图 4-2-2 单周票房统计

提交要求:

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-2-1 数据分析与可视化模块项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本	用于程序设计，每人一台。
	FTP 服务器 1 台	用于保存测试人员考试结果

工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

（3）考核时量

考核时间为 90 分钟

（4）评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-2-2 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶
	设置 rc 参数	5 分	rc 参数设置是否正确	5 分	
	读文件及筛选数据	10 分	文件和筛选数据是否获取成功	10 分	
	绘制图形	10 分	绘制图形是否符合要求	10 分	
	设置标题	10 分	设置图表标题是否符合要求	10 分	
	设置横坐标	10 分	设置图表横坐标是否符合要求	10 分	
	设置纵坐标	10 分	设置图表纵坐标是否符合	10 分	

			要求		劣影响的 本项目记0 分。
	设置图例	5分	设置图例是否符合要求	5分	
	保存和显示 图表	10分	保存和显示图表是否符合 要求	10分	
职业素 养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10 分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10 分	
总计		100分			

43. 试题编号：4-3

(1) 任务描述

2021 年影片某月的票房具体数据保存在文件 data03.csv 中。现要求你根据所提供的文件，通过 pandas 工具读取数据文件，完成相关图表的绘制。

排序, 影片名称, 单月票房 (万), 月度占比, 平均票价, 场均人次, 上映日期, 口碑指数, 月内天数

1, 送你一朵小红花, 111008, 33.3%, 37, 11, 2020-12-31, 7.8, 31
2, 拆弹专家, 261025, 18.3%, 39, 9, 2020-12-24, 7.87, 31
3, 心灵奇旅, 26207, 7.9%, 38, 10, 2020-12-25, 8.65, 31
4, 大红包, 14749, 4.4%, 33, 7, 2021-01-22, 6, 10
5, 许愿神龙, 12146, 3.6%, 35, 6, 2021-01-15, 7, 17
7, 缉魂, 10299, 3.1%, 36, 5, 2021-01-15, 7.88, 17
8, 晴雅集, 8263, 2.5%, 38, 19, 2020-12-25, 5.13, 31

图 4-3-1 文件内容

(2) 任务要求

1. 导入数据分析和可视化需用到的相关模块，其中包括完成下列①和②中要求的导入操作。

①使用 import 语句导入 matplotlib.pyplot 并取别名为 plt。

②使用 import 语句导入 pandas 并取别名为 pd。

2. 通过 pandas 中的 read_csv() 函数读取数据文件中的内容，并通过设置参数 encoding 的值为 'gbk'，实现中文的正确读取。然后筛选出绘图所需的数据列。最后利用 matplotlib 库绘制散点图。

3. 请将 pyplot 中的 rc 参数 font.sans-serif 的值设置为 "SimHei"，axes.unicode_minus 的值设成 False。

4. 散点图的标题为 "单月票房统计"。

5. 散点图纵坐标为 "票房 (万)"。

6. 散点图横坐标为影片名称。

7. 将绘图函数 scatter() 中的参数 marker 设置为 '*', 参数 color 设置为 'red'。如图 4-3-2 所示。

8. 将所绘制的散点图利用 savefig() 函数保存到与源代码相同的目录下，文件名为 "fig03.png"。

9. 使用 show() 函数显示上述绘制的图表。

10. 示例如下。

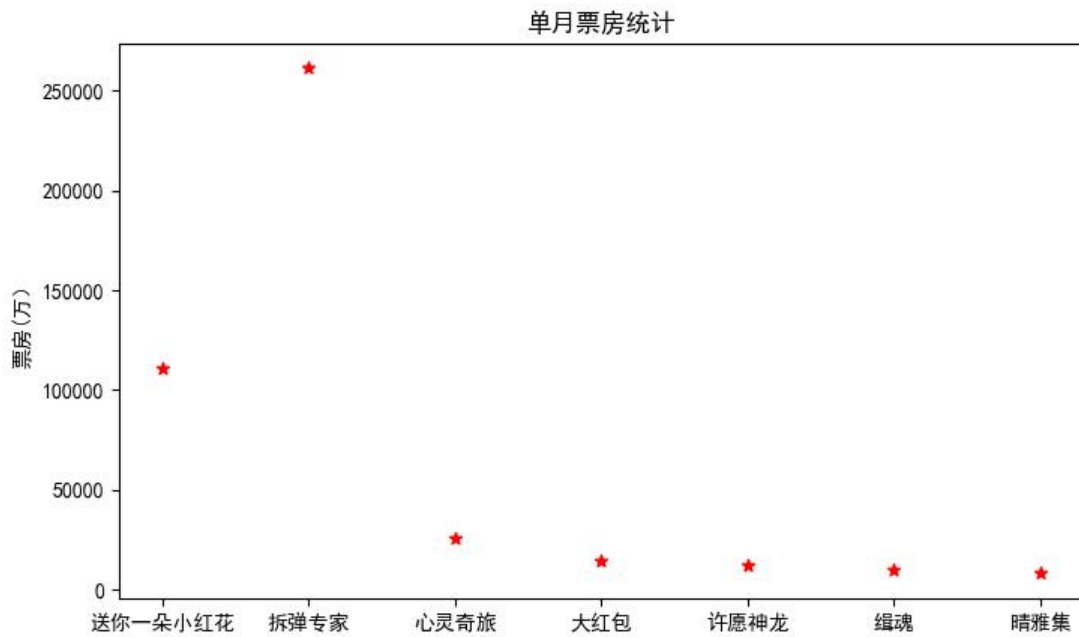


图 4-3-2 单月票房统计

提交要求:

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-3-1 数据分析与可视化模块项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	

测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-3-2 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶劣影响的本项目记 0
	设置 rc 参数	5 分	rc 参数设置是否正确	5 分	
	读文件及筛选数据	10 分	文件和筛选数据是否获取成功	10 分	
	绘制图形	15 分	绘制图形是否符合要求	15 分	
	设置标题	10 分	设置图表标题是否符合要求	10 分	
	设置横坐标	10 分	设置图表横坐标是否符合要求	10 分	
	设置纵坐标	10 分	设置图表纵坐标是否符合要求	10 分	
	保存和显示	10 分	保存和显示图表是否符合	10 分	

	图表		要求		分。
职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

44. 试题编号：4-4

(1) 任务描述

档期总票房排名具体数据保存在文件 data04.csv 中。现要求你根据所提供的文件，通过 pandas 工具读取数据文件，完成相关图表的绘制。

```
排序, 档期名称, 日期, 档期票房 (万), 总场次, 总人次 (万), 头名影片, 头名票房 (万)
1, 2021春节档, 2021年02月11日-02月17日, 778313, 2840000, 15917, 唐人街探案3, 354497
2, 2019春节档, 2019年02月04日-02月10日, 582641, 2879589, 13047, 流浪地球, 200452
3, 2018春节档, 2018年02月15日-02月21日, 572295, 2308305, 14394, 唐人街探案2, 191042
4, 2019国庆档, 2019年10月01日-10月07日, 437359, 2499600, 11667, 我和我的祖国, 191748
5, 2017春节档, 2017年01月27日-02月02日, 336913, 1846000, 8894, 西游伏妖篇, 116108
6, 2016春节档, 2016年02月07日-02月13日, 304302, 1416000, 8344, 美人鱼, 148470
7, 2018国庆档, 2018年10月01日-10月07日, 188840, 2400000, 5348, 无双, 62424
8, 2015国庆档, 2015年10月01日-10月07日, 185516, 1236688, 5660, 夏洛特烦恼, 55848
9, 2015春节档, 2015年02月18日-02月24日, 179749, 982654, 4580, 天将雄师, 45767
10, 2016国庆档, 2016年10月01日-10月07日, 158755, 1637036, 5112, 湄公河行动, 53163
```

图 4-4-1 文件内容

(2) 任务要求

1. 导入数据分析和可视化需用到的相关模块，其中包括完成下列①和②中要求的导入操作。

①使用 import 语句导入 matplotlib.pyplot 并取别名为 plt。

②使用 import 语句导入 pandas 并取别名为 pd。

2. 通过 pandas 中的 read_csv() 函数读取数据文件中的内容，并通过设置参数 encoding 的值为 'UTF-8'，实现中文的正确读取。然后筛选出绘图所需的数据列。最后利用 matplotlib 库绘制折线图。（图表的颜色可以采用默认值）。

3. 请将 pyplot 中的 rc 参数 font.sans-serif 的值设置为 "SimHei"，axes.unicode_minus 的值设成 False。

4. 折线图的标题为“档期总票房统计”。

5. 折线图纵坐标为“票房（万）”。

6. 折线图横坐标为档期名称。

7. 调用 legend() 函数，在图表的右上角显示图例，如图 4-4-2 所示。

8. 将所绘制的折线图利用 savefig() 函数到与源代码相同的目录下，文件名为“fig04.png”。

9. 使用 show() 函数显示上述绘制的图表。

10. 示例如下。

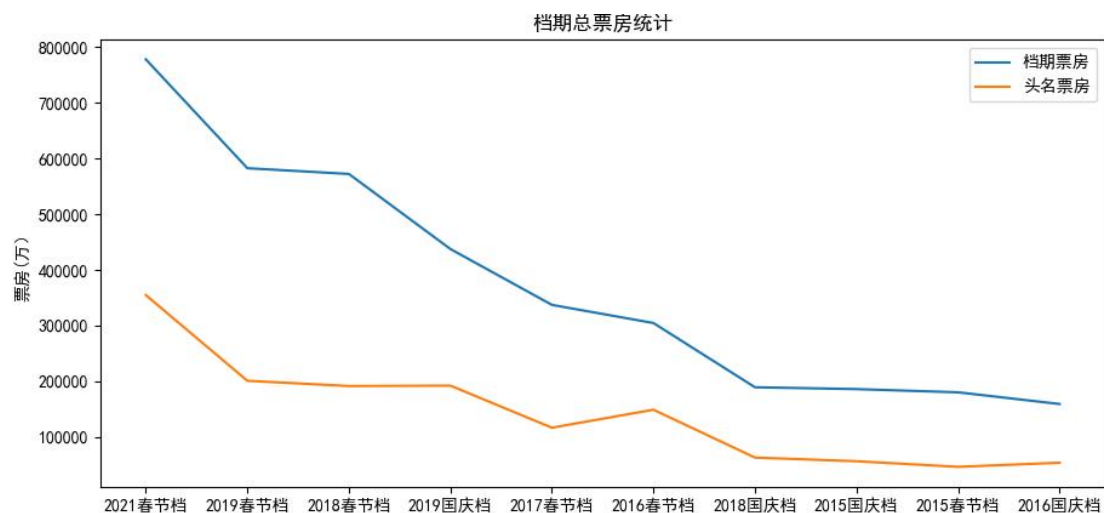


图 4-4-2 档期总票房

提交要求:

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-4-1 数据分析与可视化模块项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本	用于程序设计，每人一台。
	FTP 服务器 1 台	用于保存测试人员考试结果

工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	
测评专家	现场测评专家：	在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	测评专家满足任一条件
	结果测评专家：	在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-4-2 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪
	设置 rc 参数	5 分	rc 参数设置是否正确	5 分	
	读文件及筛选数据	10 分	文件和筛选数据是否获取成功	10 分	
	绘制图形	10 分	绘制图形是否符合要求	10 分	
	设置标题	10 分	设置图表标题是否符合要求	10 分	
	设置横坐标	10 分	设置图表横坐标是否符合要求	10 分	

	设置纵坐标	10分	设置图表纵坐标是否符合要求	10分	律、造成恶劣影响的 本项目记0分。
	设置图例	5分	设置图例是否符合要求	5分	
	保存和显示图表	10分	保存和显示图表是否符合要求	10分	
职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

45. 试题编号：4-5

(1) 任务描述

内地总票房排名具体数据保存在文件 data5.csv 中。现要求你根据所提供的
数据文件，通过 pandas 工具读取 data5.csv 文件，完成相关图表的绘制。

```
排序, 影片名称, 类型, 总票房 (万), 平均票价, 场均人次, 国家及地区, 上映日期  
1, 战狼2, 动作, 568832, 36, 38, 中国, 2017-07-27  
2, 你好, 李焕英, 喜剧, 541330, 45, 24, 中国, 2021-02-12  
3, 哪吒之魔童降世, 动画, 503502, 36, 23, 中国, 2019-07-26  
4, 流浪地球, 科幻, 468680, 45, 29, 中国, 2019-02-05  
5, 唐人街探案3, 喜剧, 452234, 48, 29, 中国, 2021-02-12
```

图 4-5-1 文件内容

(2) 任务要求

1. 导入数据分析和可视化需用到的相关模块，其中包括完成下列①和②中要求的导入操作。

①使用 import 语句导入 matplotlib.pyplot 并取别名为 plt。

②使用 import 语句导入 pandas 并取别名为 pd。

2. 通过 pandas 中的 read_csv() 函数读取数据文件中的内容，并通过设置参数 encoding 的值为 'gbk'，实现中文的正确读取。然后筛选出绘图所需的数据列。

最后利用 matplotlib 库绘制饼图。（图表的颜色可以采用默认值）。

3. 请将 pyplot 中的 rc 参数 font.sans-serif 的值设置为 “SimHei”，axes.unicode_minus 的值设成 False。

4. 饼图的标题为 “内地总票房统计”。

5. 将绘图函数 pie() 中的参数 labels 设置为 ‘影片名称’，参数 autopct 设置为 ‘%.1f%%’，如图 4-5-2 所示。

6. 调用 legend() 函数，在图表的右上角显示图例。

7. 将所绘制的饼图利用 savefig() 函数保存到与源代码相同的目录下，文件名为 “fig05.png”。

8. 使用 show() 函数显示上述绘制的图表。

9. 示例如下。

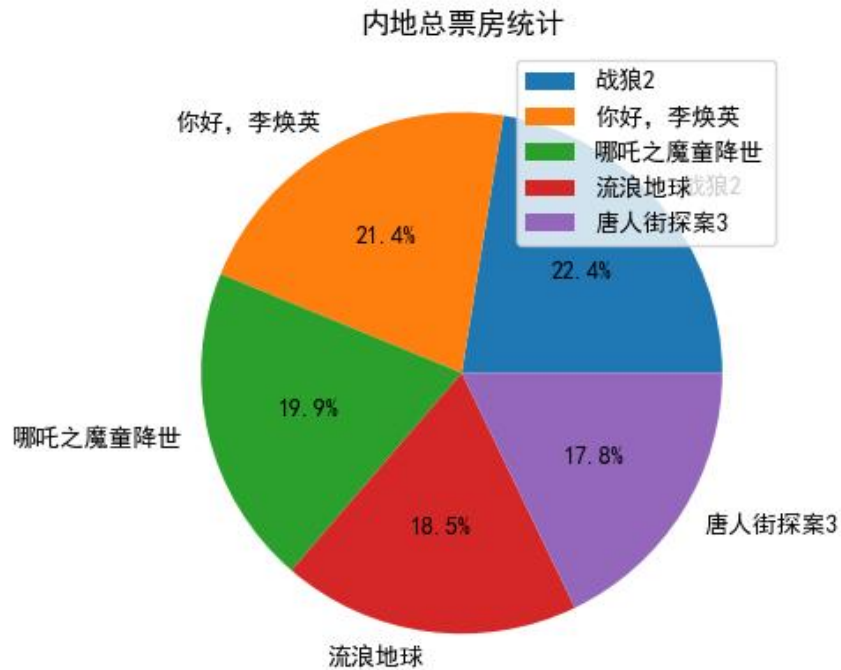


图 4-5-2 内地总票房

提交要求:

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-5-1 数据分析与可视化模块项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	

测 评 专 家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。	测评专家满足任一条件	
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-5-2 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任 务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪律、造成恶
	设置 rc 参数	5 分	rc 参数设置是否正确	5 分	
	读文件及筛选数据	10 分	文件和筛选数据是否获取成功	10 分	
	绘制图形	25 分	绘制图形是否符合要求	25 分	
	设置标题	10 分	设置图表标题是否符合要求	10 分	
	设置图例	10 分	设置图例是否符合要求	10 分	
	保存和显示图表	10 分	保存和显示图表是否符合要求	10 分	

职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	劣影响的 本项目记0分。
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			

项目 2: pyecharts 数据可视化

46. 试题编号: 4-6

(1) 任务描述

2021 年 3 月上旬长沙市空气质量统计历史数据保存在表 4-6-1 中。现要求你根据表中的数据, 完成相关图表的绘制。

表 4-6-1 空气质量统计数据

日期	AQI	质量等级	PM2.5	PM10	S02	CO	N02	O3_8h
03-01	26	优	14	9	5	0.9	20	52
03-02	38	优	26	22	6	0.7	23	60
03-03	55	良	39	28	6	0.8	32	54
03-04	49	优	34	32	7	1.1	32	49
03-05	40	优	22	19	5	1	32	33
03-06	46	优	32	25	5	0.8	20	41
03-07	60	良	43	38	5	0.7	21	20
03-08	63	良	45	33	5	0.8	29	13
03-09	65	良	47	32	6	0.8	31	30
03-10	57	良	40	24	6	0.8	36	16

字段说明:

AQI: 空气质量指数;

PM2.5: 细颗粒物粒径小于等于 2.5 微米;

PM10: 细颗粒物粒径小于等于 10 微米;

S02: 二氧化硫平均浓度值;

CO: 一氧化碳平均浓度值;

N02: 二氧化氮平均浓度值;

O3_8h: 臭氧 8 小时平均浓度值

(2) 任务要求

1. 导入绘图需用到的相关模块, 其中包括完成下列①和②中要求的导入操作。

①使用 from import 语句导入 pyecharts.charts 中的折线图 Line。

②使用 from import 语句导入 pyecharts 中的 options 并取别名为 opts。

2. 将绘图需要的数据使用列表保存，然后利用 pyecharts 库，绘制折线图。（图的颜色采用默认值）。
3. 调用 add_xaxis() 函数，设置 x 轴的数据为表 4-6-1 中的日期列。
4. 调用 add_yaxis() 函数，设置 y 轴的数据为表 4-6-1 中的 AQI 列。
5. 实现图表的参数配置：调用 set_global_opts() 函数，将其参数 title_opts 的值设置为 opts.TitleOpts(title="空气质量指数")，实现给折线图添加标题，如图 4-6-1 所示。
6. 将所绘制的折线图利用 render() 函数保存到与源代码相同目录下，其中图表文件名为“fig06.html”。
7. 示例如下。

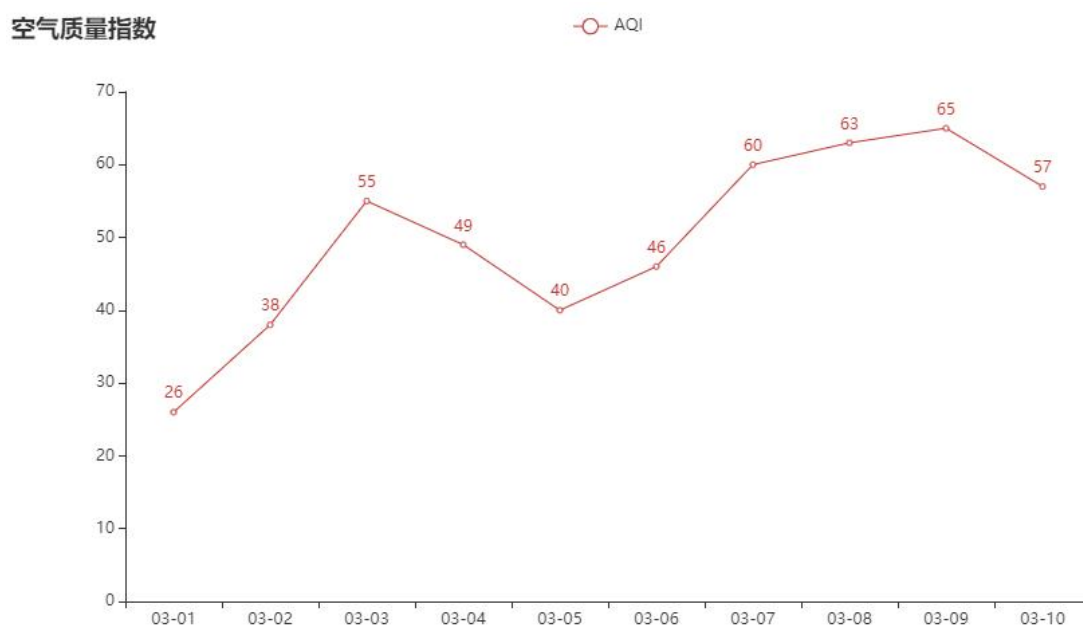


图 4-6-1 空气质量指数

提交要求:

- 1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。
- 2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-6-2 数据分析与可视化模块项目实施条件

项目	基本实施条件	备注
----	--------	----

场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-6-3 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、
	保存数据	10 分	数据的保存是否符合要求	10 分	

	绘制图形	20分	绘制图形是否符合要求	20分	没有按要求填写相关信息,本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	设置标题	10分	设置图表标题是否符合要求	10分	
	设置横坐标	10分	图表横坐标是否符合要求	10分	
	设置纵坐标	10分	图表纵坐标是否符合要求	10分	
	保存图表	10分	保存图表是否符合要求	10分	
职业素养	专业素养	10分	代码符合代码开发规范,命名规范,能做到见名知意;缩进统一,方便阅读;注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10分	
总计		100分			

47. 试题编号：4-7

(1) 任务描述

2021年6月上旬武汉空气质量统计历史数据保存在表4-7-1中。现要求你根据表中的数据，完成相关图表的绘制。

表4-7-1 空气质量统计数据

日期	AQI	质量等级	PM2.5	PM10	S02	CO	N02	O3_8h
06-01	101	轻度污染	33	65	9	0.8	45	161
06-02	119	轻度污染	38	74	7	0.9	39	180
06-03	58	良	20	41	5	0.7	25	109
06-04	75	良	17	41	7	0.6	33	130
06-05	95	良	20	50	8	0.6	41	154
06-06	129	轻度污染	26	60	10	0.8	40	191
06-07	102	轻度污染	26	49	11	0.8	33	162
06-08	134	轻度污染	36	64	12	1	48	197
06-09	125	轻度污染	35	65	11	1	47	187
06-10	53	良	33	51	6	0.9	40	103

字段说明：

AQI: 空气质量指数；

PM2.5: 细颗粒物粒径小于等于2.5微米；

PM10: 细颗粒物粒径小于等于10微米；

S02: 二氧化硫平均浓度值；

CO: 一氧化碳平均浓度值；

N02: 二氧化氮平均浓度值；

O3_8h: 臭氧8小时平均浓度值

(2) 任务要求

1. 导入绘图需用到的相关模块，其中包括完成下列①和②中要求的导入操作。

- ①使用 `from import` 语句导入 `pyecharts.charts` 中的柱状图 `Bar`。
- ②使用 `from import` 语句导入 `pyecharts` 中的 `options` 并取别名为 `opts`。
2. 将绘图需要的数据使用列表保存，然后利用 `pyecharts` 库，绘制柱状图。（图的颜色采用默认值）。
3. 调用 `add_xaxis()` 函数，设置 x 轴的数据为表 4-7-1 中的日期列。
4. 调用 `add_yaxis()` 函数，设置 y 轴的数据为表 4-7-1 中的 `NO2` 列。
5. 实现图表的参数配置：调用 `set_global_opts()` 函数，将其参数 `title_opts` 的值设置为 `opts.TitleOpts(title="空气质量 NO2")`，实现给柱状图添加标题，如图 4-7-1 所示。
6. 将所绘制的柱状图利用 `render()` 函数保存到与源代码相同目录下，其中图表文件名为“`fig07.html`”。
7. 示例如下。

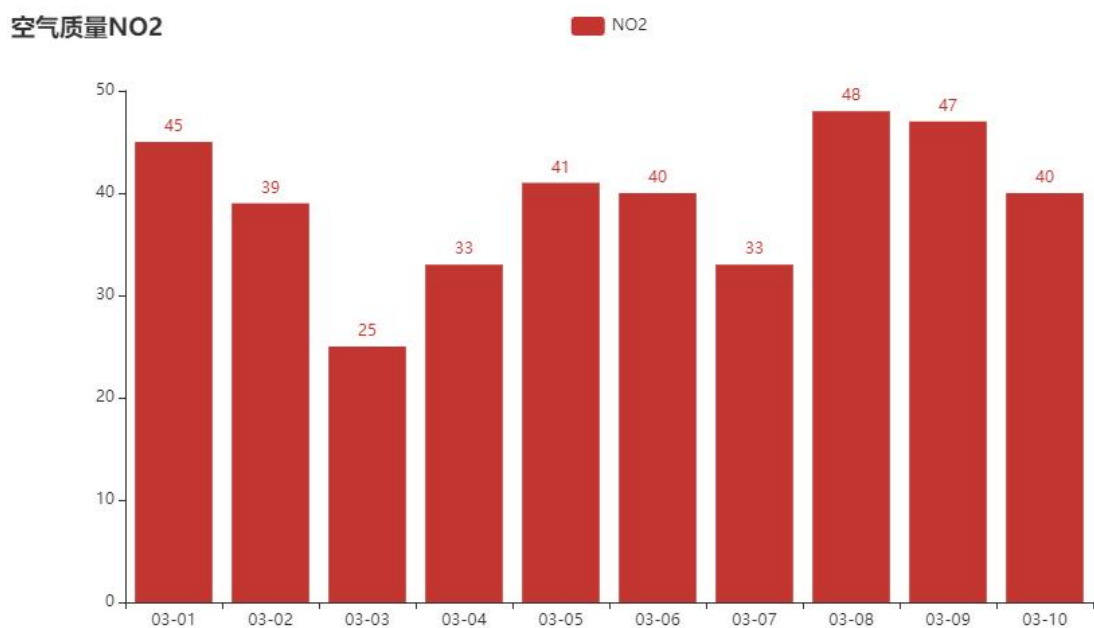


图 4-7-1 空气质量 NO2

提交要求:

- 1) 在“`e:\技能抽查提交资料\`”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。
- 2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-7-2 数据分析与可视化模块项目实施条件

项目	基本实施条件		备注
场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-7-3 数据分析与可视化模块考核评价标准

评价内容	配分	评分标准	备注
------	----	------	----

工作任务	导入相关库	10分	导入相关库是否正确	10分	1、考试舞弊、抄袭、没有按要求填写相关信息,本项目记0分。
	保存数据	10分	数据的保存是否符合要求	10分	
	绘制图形	20分	绘制图形是否符合要求	20分	
	设置标题	10分	设置图表标题是否符合要求	10分	
	设置横坐标	10分	图表横坐标是否符合要求	10分	
	设置纵坐标	10分	图表纵坐标是否符合要求	10分	
	保存图表	10分	保存图表是否符合要求	10分	
职业素养	专业素养	10分	代码符合代码开发规范,命名规范,能做到见名知意;缩进统一,方便阅读;注释规范。	0-10分	2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	道德规范	10分	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10分	
总计		100分			

48. 试题编号：4-8

(1) 任务描述

部分省份的空气质量统计历史数据保存在表 4-8-1 中。现要求你根据表中的数据，完成相关图表的绘制。

表 4-8-1 空气质量统计数据

序号	城市	省份	AQI	质量等级	PM2.5	PM10
1	普洱	云南	21	优	9	17
2	三亚	海南	23	优	7	17
3	德宏州	云南	25	优	11	20
4	海口	海南	25	优	8	19
5	黑河	黑龙江	25	优	8	19
6	临沧	云南	26	优	15	18
7	伊春	黑龙江	27	优	9	15
8	保山	云南	29	优	9	15

字段说明：

AQI：空气质量指数；

PM2.5：细颗粒物粒径小于等于 2.5 微米；

PM10：细颗粒物粒径小于等于 10 微米；

(2) 任务要求

1. 导入绘图需用到的相关模块，其中包括完成下列①和②中要求的导入操作。

①使用 `from import` 语句导入 `pyecharts.charts` 中的散点图 `Scatter`。

②使用 `from import` 语句导入 `pyecharts` 中的 `options` 并取别名为 `opts`。

2. 将绘图需要的数据使用列表保存，然后利用 `pyecharts` 库，绘制散点图。（图的颜色采用默认值）。

3. 调用 `add_xaxis()` 函数，设置 x 轴的数据为表 4-8-1 中的城市列。

4. 调用 `add_yaxis()` 函数，设置 y 轴的数据为表 4-8-1 中的 PM10 列。

5. 实现图表的参数配置：调用 `set_global_opts()` 函数，将其参数 `title_opts` 的值设置为 `opts.TitleOpts(title="空气质量 PM10")`，实现给散点图添加标题，如图 4-8-1 所示。

6. 将所绘制的散点图利用 `render()` 函数保存到与源代码相同目录下，其中图表文件名为“`fig08.html`”。

7. 示例如下。

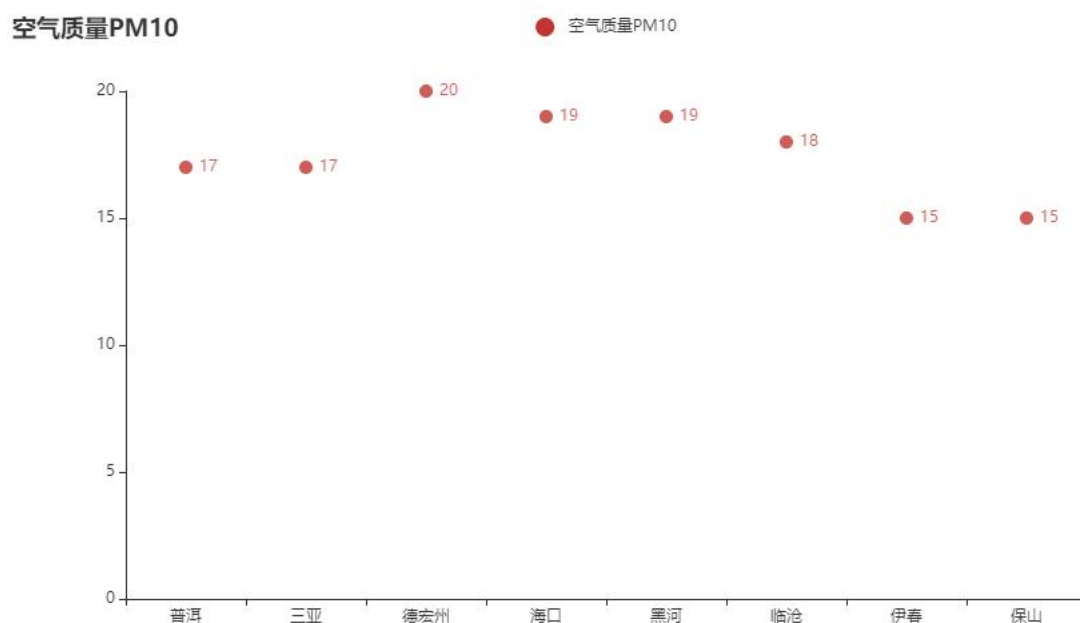


图 4-8-1 空气质量 PM10

提交要求:

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-8-2 数据分析与可视化模块项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本	用于程序设计，每人一台。
	FTP 服务器 1 台	用于保存测试人员考试结果

工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-8-3 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。 2、严重违反考场纪
	保存数据	10 分	数据的保存是否符合要求	10 分	
	绘制图形	20 分	绘制图形是否符合要求	20 分	
	设置标题	10 分	设置图表标题是否符合要求	10 分	
	设置横坐标	10 分	图表横坐标是否符合要求	10 分	
	设置纵坐标	10 分	图表纵坐标是否符合要求	10 分	
	保存图表	10 分	保存图表是否符合要求	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范，命	0-10	

养			名规范，能做到见名知意； 缩进统一，方便阅读；注释 规范。	分	律、造成恶劣影响的 本项目记0 分。
	道德规范	10分	着装干净、整洁。举止文明， 遵守考场纪律，按顺序进出 考场。	0-10 分	
总计		100分			

49. 试题编号：4-9

(1) 任务描述

2021年3月上旬长沙市空气质量统计历史数据保存在表4-9-1中。现要求你根据表中的数据，完成相关图表的绘制。

表4-9-1 空气质量统计数据

序号	省份	城市数	AQI	质量等级	PM2.5	PM10
1	海南	2	24	优	7	18
2	云南	16	31	优	11	20
3	黑龙江	13	34	优	10	22
4	贵州	9	37	优	10	20
5	广西	14	40	优	15	31
6	吉林	9	42	优	12	26
7	福建	9	43	优	14	30
8	广东	21	47	优	14	27
9	西藏	7	47	优	7	18
10	湖南	14	50	优	16	30

字段说明：

AQI：空气质量指数；

PM2.5：细颗粒物粒径小于等于2.5微米；

PM10：细颗粒物粒径小于等于10微米；

(2) 任务要求

1. 导入绘图需用到的相关模块，其中包括完成下列①和②中要求的导入操作。

①使用 `from import` 语句导入 `pyecharts.charts` 中的柱状图 `Bar`。

②使用 `from import` 语句导入 `pyecharts` 中的 `options` 并取别名为 `opts`。

2. 将绘图需要的数据使用列表保存，然后利用 `pyecharts` 库，绘制横向柱状图。

（图表的颜色采用默认值）。

3. 调用 `add_xaxis()` 函数，设置 x 轴的数据为表4-9-1中的省份列。

4. 调用 `add_yaxis()` 函数，设置 y 轴的数据为表4-9-1中的PM2.5列。

5. 实现图表的参数配置：调用 `set_global_opts()` 函数，将其参数 `title_opts` 的值设置为 `opts.TitleOpts(title="空气质量PM2.5")`，实现给横向柱状图添加

标题。

6. 实现图表的参数配置：调用 `set_series_opts()` 函数，将其参数 `label_opts` 的值设置为 `opts.LabelOpts(position="right")`，实现将数据标签在条形的右侧显示，如图 4-9-1 所示。

7. 实现图表的参数配置：调用 `reversal_axis()` 函数，实现柱状图横轴、纵轴的交流。

8. 将所绘制的横向柱状图利用 `render()` 函数保存到与源代码相同目录下，其中图表文件名为“`fig09.html`”。

9. 示例如下。

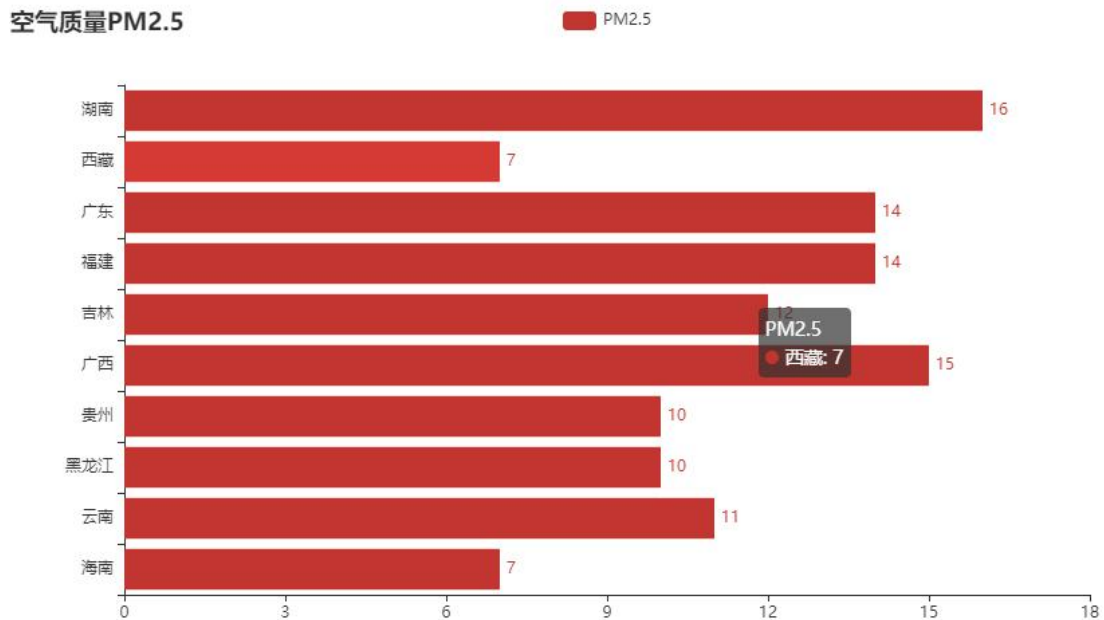


图 4-9-1 空气质量 PM2.5

提交要求：

1) 在“`e:\技能抽查提交资料\`”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-9-2 数据分析与可视化模块项目实施条件

项目	基本实施条件	备注
----	--------	----

场地	能同时容纳 30 人以上现场考核		
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本		用于程序设计，每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts1.9.0、pyecharts_snapshot）	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-9-3 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、
	保存数据	10 分	数据的保存是否符合要求	10 分	

	绘制图形	20分	绘制图形是否符合要求	20分	没有按要求填写相关信息,本项目记0分。 2、严重违反考场纪律、造成恶劣影响的本项目记0分。
	设置标题	10分	设置图表标题是否符合要求	10分	
	设置横坐标	10分	图表横坐标是否符合要求	10分	
	设置纵坐标	10分	图表纵坐标是否符合要求	10分	
	保存图表	10分	保存图表是否符合要求	10分	
职业素养	专业素养	10分	代码符合代码开发规范,命名规范,能做到见名知意;缩进统一,方便阅读;注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明,遵守考场纪律,按顺序进出考场。	0-10分	
总计		100分			

50. 试题编号：4-10

(1) 任务描述

2021 年长沙市某天 0 点—12 点空气质量统计历史数据保存在表 4-10-1 中。现要求你根据表中的数据，完成相关图表的绘制。

表 4-10-1 空气质量统计数据

时间	AQI	PM2.5
0 点	49	25
1 点	49	28
2 点	47	27
3 点	46	26
4 点	43	25
5 点	40	25
6 点	37	24
7 点	36	23
8 点	37	23
9 点	39	22
10 点	37	23
11 点	36	24
12 点	39	26

字段说明：

AQI：空气质量指数；

PM2.5：细颗粒物粒径小于等于 2.5 微米；

(2) 任务要求

1. 导入绘图需用到的相关模块，其中包括完成下列①和②中要求的导入操作。
 - ①使用 `from import` 语句导入 `pyecharts.charts` 中的柱状图 `Bar`。
 - ②使用 `from import` 语句导入 `pyecharts` 中的 `options` 并取别名为 `opts`。
2. 将绘图需要的数据使用列表保存，然后利用 `pyecharts` 库，绘制双柱状图。（图的颜色采用默认值）。
3. 调用 `add_xaxis()` 函数，设置 x 轴的数据为表 4-10-1 中的时间列。
4. 调用 `add_yaxis()` 函数，设置 y 轴的数据为表 4-10-1 中的 AQI 和 PM2.5 列。
5. 实现图表的参数配置：调用 `set_global_opts()` 函数，将其参数 `title_opts` 的值设置为 `opts.TitleOpts(title="空气质量指数 AQI 和 PM2.5")`，实现给双柱

状图添加标题。

6. 将所绘制的双柱状图利用 render() 函数保存到与源代码相同目录下，其中图表文件名为“fig10.html”。

7. 示例如下。

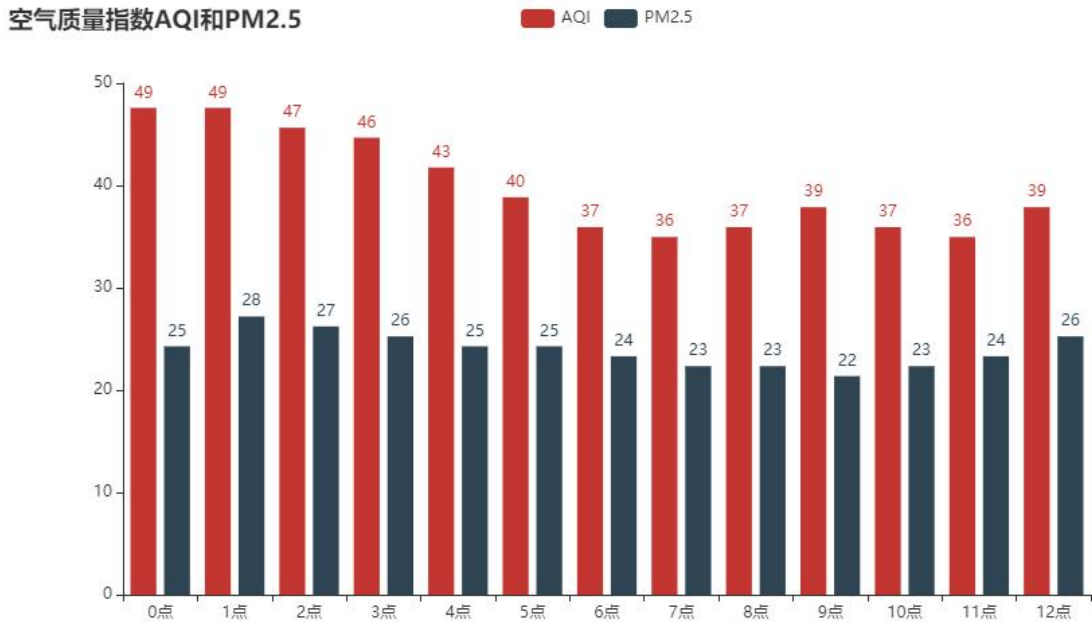


图 4-10-1 空气质量指数 AQI 和 PM2.5

提交要求:

1) 在“e:\技能抽查提交资料\”文件夹内创建考生文件夹，考生文件夹的命名规则：考生学校+考生号+考生姓名，示例：永州职业技术学院 01 张三。

2) 考生文件夹内保存代码源文件及试题文件（填写代码+运行截图），代码源文件以“姓名_题号.py”命名，最终将考生文件夹进行压缩后提交。

表 4-10-2 数据分析与可视化模块项目实施条件

项目	基本实施条件	备注
场地	能同时容纳 30 人以上现场考核	
设备	30 台以上的主流计算机 安装 Windows 7 或更高版本	用于程序设计，每人一台。
	FTP 服务器 1 台	用于保存测

			试人员考试结果
工具	开发工具	Pycharm2019 或更高版本（安装库：matplotlib、pandas、pyecharts）	
测评专家	现场测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		测评专家满足任一条件
	结果测评专家：在本行业具有 3 年以上的从业经验（工程师及以上职称）或从事本专业具有 5 年以上的教学经验（副高及以上职称），或具有软件设计师、系统分析师、数据库设计师资格证书（2 人/场）。		

(3) 考核时量

考核时间为 90 分钟

(4) 评分标准

数据分析与可视化模块的考核实行 100 分制，评价内容包括职业素养、工作任务完成情况两个方面。其中，职业素养占该项目总分的 20%，工作任务完成质量占该项目总分的 80%。具体评价标准见下表：

表 4-10-3 数据分析与可视化模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	导入相关库	10 分	导入相关库是否正确	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分。
	保存数据	10 分	数据的保存是否符合要求	10 分	
	绘制图形	20 分	绘制图形是否符合要求	20 分	
	设置标题	10 分	设置图表标题是否符合要求	10 分	
	设置横坐标	10 分	图表横坐标是否符合要求	10 分	
	设置纵坐标	10 分	图表纵坐标是否符合要求	10 分	

	保存图	10分	保存图表是否符合要求	10分	2、严重违反考场纪律、造成恶劣影响的本项目记0分。
职业素养	专业素养	10分	代码符合代码开发规范，命名规范，能做到见名知意；缩进统一，方便阅读；注释规范。	0-10分	
	道德规范	10分	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。	0-10分	
总计		100分			