网络空间安全学院
School of Cyberspace Security

湖南信息职业技术学院
学生专业技能考核题库

专业：　大数据技术

部门：　网络空间安全学院

编制：　云计算与大数据教研室

2021 年 7 月

湖南信息职业技术学院

大数据技术专业学生专业技能考核题库

本专业技能考核以《湖南信息职业技术学院大数据技术专业学生专业技能考核标准》为编制依据，通过设置大数据开发基础模块、数据采集模块、数据清洗与挖掘应用模块、数据分析与可视化模块等 4 个技能考核模块，测试学生的编程能力、数据采集能力、数据清洗能力、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。

大数据开发基础和大数据平台部署与开发模块以企、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，基本涵盖了运维工程师、数据可视化工程师和数据分析师等岗位从事项目设计与开发工作所需的基本技能。

数据采集模块主要以大数据采集技术和Python工具为背景，通过网络爬虫实现对具体页面的数据进行采集。本模块基本涵盖了数据分析师岗位所需的数据采集核心技能。

数据清洗与挖掘应用模块主要以大数据清洗技术和Spark等工具为背景，将实际应用中采集到的数据进行清洗、挖掘与存储。本模块基本涵盖了数据分析工程师岗位所需的数据清洗核心技能。

数据分析与可视化模块主要以大数据可视化技术和工具为背景，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所需的数据展示核心技能。

目 录

一、 专业基本技能 1
模块一 大数据开发基础 1
项目 1：大数据编程基础 1
1.试题编号：1-1-1，控制台打印 9*9 乘法口诀表，任意输入三个整数排序后输出 1
2.试题编号：1-1-2，输入学生成绩返回等级，输入两个正整数求最大公约数和最小公倍数 4
3.试题编号：1-1-3，猴子吃桃问题，兔子问题 7
4.试题编号：1-1-4，找 1000 以内的所有“水仙花数”，自由落体问题 10
5.试题编号：1-1-5，求 100 以内所有偶数的和，企业发放奖金问题 12
项目 2：大数据平台部署与开发 15
6.试题编号：1-2-1，Hadoop 平台完全分布式部署； 15
7.试题编号：1-2-2，Hadoop 常用命令—创建目录； 19
8.试题编号：1-2-3，Hadoop 常用命令—上传文件； 23
9.试题编号：1-2-4，Hadoop 常用命令—下载文件； 27
10.试题编号：1-2-5，Hadoop 常用命令—删除文件/目录； 32
11.试题编号：1-2-6，使用 MapReduce 程序计算学生期末考试各科最高分； 37
12.试题编号：1-2-7，使用 MapReduce 程序计算学生期末考试各科最低分； 43
13.试题编号：1-2-8，使用 MapReduce 程序计算学生期末考试各科平均分； 49
14.试题编号：1-2-9，使用 MapReduce 程序计算学生期末考试各科区间分布情况； 55
15.试题编号：1-2-10，使用 MapReduce 程序计算学生期末考试各科成绩进行排序； 61
二、岗位核心技能 67
模块二 数据采集 67
项目 1：基于 Flume 的数据采集 67
16.试题编号：2-1-1，使用 Flume 采集指定文件数据 67
17.试题编号：2-1-2，使用 Flume 采集 shell 命令或者脚本的结果数据 72
18.试题编号：2-1-3，使用 Flume 采集通过 TCP 协议发送的数据 ... 76
19.试题编号：2-1-4，使用 Flume 采集通过运营支撑数据 81
20.试题编号：2-1-5，日记数据采集到 HDFS 的数据采集系统 86
项目 2：基于 kafka 的消息队列数据采集 92
21.试题编号：2-2-1，单 source、单 channel 构建数据采集系统 ... 92
22.试题编号：2-2-2，单 source、多 channel 构建数据采集系统 ... 97
23.试题编号：2-2-3，多 source、多 channel 构建数据采集系统 .. 103

24.试题编号：2-2-4，多 source、单 channel 构建数据采集系统 ..109
25.试题编号：2-2-5，日记数据采集到消息队列的采集系统 115
模块三 数据清洗与挖掘应用 121
项目 1：基于 kettle 的数据清洗 121
26.试题编号：3-1-1：Excel 数据清洗 121
27.试题编号：3-1-2：TXT 数据和 XML 数据清洗 126
28.试题编号：3-1-3：csv 数据清洗 131
29.试题编号：3-1-4，JS 数据清洗 138
30.试题编号：3-1-5：CSV 数据综合清洗 142
项目 2 Spark 大数据处理与分析 147
31 试题编号：3-2-1：Spark 大数据分析平台搭建 147
32 试题编号：3-2-2：Spark 大数据分析平台搭建 152
33 试题编号：3-2-3：Spark 开发-网站用户访问日志数据分析 156
34 试题编号：3-2-4：Spark 开发-网站用户访问日志数据分析 161
35 试题编号：3-2-5：Spark 开发-股票数据分析 164
36 试题编号：3-2-6：Spark 开发-股票数据分析与预测 167
37 试题编号：3-2-7：Spark 开发-词频统计 171
38 试题编号：3-2-8：Spark 开发-Apache 日志分析 173
39 试题编号：3-2-9：Spark 开发-SparkStreaming 实时网络处理数据
..... 176
40 试题编号：3-2-10：Spark 开发-SparkStreaming 实时 HDFS 处理数据
..... 179
模块四 数据分析与可视化 182
项目 1：基于 matplotlib 的数据分析和可视化 182
41.试题编号：4-1-1，单日票房数据分析和可视化 182
42.试题编号：4-1-2，单周票房数据分析和可视化 185
43.试题编号：4-1-3，单月票房数据分析和可视化 188
44.试题编号：4-1-4，档期总票房数据分析和可视化 192
45.试题编号：4-1-5，内地总票房排名数据分析和可视化 195
项目 2：基于 pyecharts 的数据可视化 199
46.试题编号：4-2-1，空气质量指数 AQI 数据可视化 199
47.试题编号：4-2-2，空气质量 NO2 数据可视化 202
48.试题编号：4-2-3，空气质量 PM10 数据可视化 205
49.试题编号：4-2-4，空气质量 PM2.5 数据可视化 208
50.试题编号：4-2-5，空气质量指数 AQI 和 PM2.5 数据可视化 212

一、专业基本技能

模块一 大数据开发基础

项目 1：大数据编程基础

1.试题编号：1-1-1，控制台打印 9*9 乘法口诀表，任意输入三个整数排序后输出

（1）任务描述

任务一 控制台打印 9*9 乘法口诀表（40 分）

任务要求：

编程实现控制台打印 9*9 乘法口诀。请分行与列考虑，共 9 行 9 列，i 控制行，j 控制列。

1.根据题目描述，编程程序。

2.程序执行示意图如下。

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

Process finished with exit code 0
```

图 1-1 输出 9*9 口诀图

任务二 任意输入三个整数排序后输出（40 分）

任务要求：

输入三个整数 x，y，z，请把这三个数由小到大输出。

任务分析，我们想办法把最小的数放到 x 上，先将 x 与 y 进行比较，如果 x>y 则将 x 与 y 的值进行交换，然后再用 x 与 z 进行比较，如果 x>z 则将 x 与 z 的值进行交换，这样能使 x 最小。

1.根据任务描述与分析，编写程序。

2.程序执行结果如图所示。

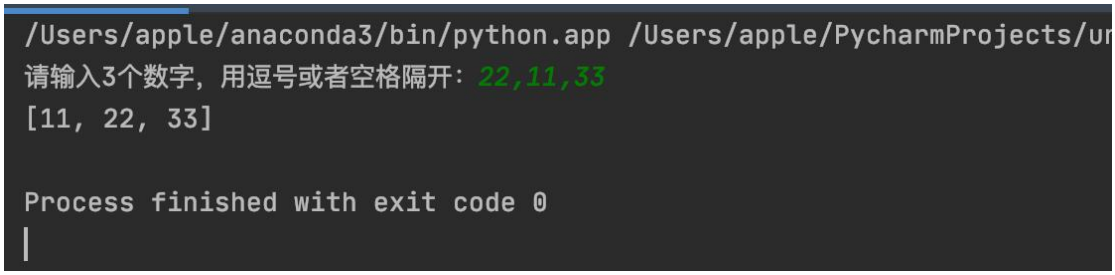


图 1-2 从小到大输出三个整数

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

表 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 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 120 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 考核评价标准

<table>
<tr><th colspan="2">评价内容</th><th>分值</th><th colspan="2">评分标准</th><th>备注</th></tr>
<tr><td rowspan="5">工作任务一</td><td rowspan="3">9*9口诀</td><td rowspan="3">30分</td><td>输出的9*9口诀嵌套循环符合要求</td><td>10分</td><td rowspan="12">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。</td></tr>
<tr><td>循环次数设置符合要求</td><td>10分</td></tr>
<tr><td>输出格式顺序及形式符合要求</td><td>10分</td></tr>
<tr><td>变量使用</td><td>5分</td><td>变量声明是否符合要求</td><td>5分</td></tr>
<tr><td>空格</td><td>5分</td><td>换行及空格是否符合要求</td><td>5分</td></tr>
<tr><td rowspan="5">工作任务二</td><td rowspan="3">对输入的数据是否排序</td><td rowspan="3">30分</td><td>比较大小是否符合逻辑</td><td>10分</td></tr>
<tr><td>变量转换是否正确</td><td>10分</td></tr>
<tr><td>返回结果形式是否正确</td><td>10分</td></tr>
<tr><td>变量使用</td><td>5分</td><td>变量声明是否符合要求</td><td>5分</td></tr>
<tr><td>循环和排序</td><td>5分</td><td>排序是否符合要求</td><td>5分</td></tr>
<tr><td rowspan="2">职业素养</td><td>专业素养</td><td>10分</td><td>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。</td><td>0-10分</td></tr>
<tr><td>道德规范</td><td>10分</td><td>着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。</td><td>0-10分</td></tr>
<tr><td colspan="2">总计</td><td colspan="4">100分</td></tr>
</table>

2.试题编号：1-1-2，输入学生成绩返回等级，输入两个正整数求最大公约数和最小公倍数

（1）任务描述

任务一 输入学生成绩返回等级

任务要求：

从键盘接收 100 分制成绩（1-100），要求输出其对应的成绩等级 A-E。

其中学习成绩>=90 分的同学用 A 表示，80-89 分之间用 B 表示，70-79 分之间的用 C 表示，60-69 分之间的用 D 表示，60 分以下为 E。利用条件运算符的嵌套来完成此题。

1.根据任务描述与分析，编写程序。

2.程序执行结果如图所示。

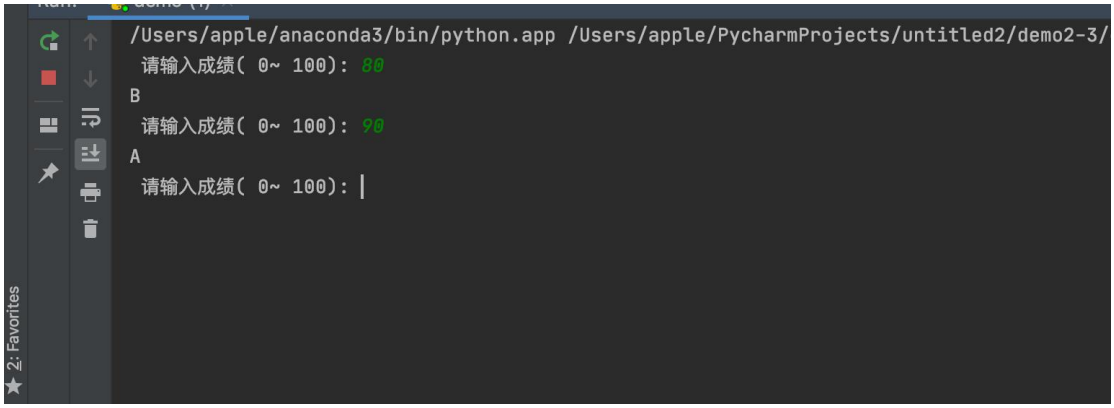


图 1-3 学生成绩分类结果图

任务二 输入两个正整数求最大公约数和最小公倍数

任务要求：

程序风格良好(使用自定义注释模板)，辗转相除法算法解决最大公约数问题，提供友好的输入输出。

1. 根据任务分析程序，分析辗转相除法算法执行流程。

2. 正确调试程序。

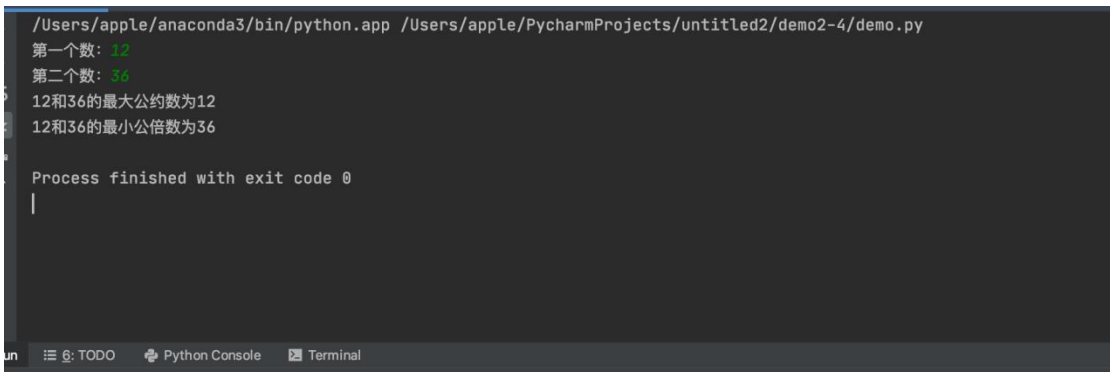


图 1-4　最大公约数和最小公倍数输出图

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

表 1-3 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 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	
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 120 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 考核评价标准

<table>
<tr><th colspan="2">评价内容</th><th>配分</th><th colspan="2">评分标准</th><th>备注</th></tr>
<tr><td rowspan="5">工作任务一</td><td rowspan="3">对输入的成绩是否分类</td><td rowspan="3">30 分</td><td>正确使用多分支判定语句</td><td>10 分</td><td rowspan="12">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。</td></tr>
<tr><td>成绩的判定范围正确</td><td>10 分</td></tr>
<tr><td>返回结果区间正确</td><td>10 分</td></tr>
<tr><td>变量使用</td><td>5 分</td><td>变量声明是否符合要求</td><td>5 分</td></tr>
<tr><td>判断语句</td><td>5 分</td><td>判断是否符合要求</td><td>5 分</td></tr>
<tr><td rowspan="5">工作任务二</td><td rowspan="3">最大公约数、最小公倍数计算</td><td rowspan="3">30 分</td><td>正确定义变量</td><td>10 分</td></tr>
<tr><td>正确执行逻辑判定语句</td><td>12</td></tr>
<tr><td>结果输出正确</td><td>8 分</td></tr>
<tr><td>辗转相除法</td><td>5 分</td><td>辗转相除法</td><td>5 分</td></tr>
<tr><td>判断、循环语句</td><td>5 分</td><td>判断是否符合要求</td><td>5 分</td></tr>
<tr><td rowspan="2">职业素养</td><td>专业素养</td><td>10 分</td><td>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。</td><td>0-10 分</td></tr>
<tr><td>道德规范</td><td>10 分</td><td>着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。</td><td>0-10 分</td></tr>
<tr><td colspan="2">总计</td><td colspan="4">100 分</td></tr>
</table>

3.试题编号：1-1-3，猴子吃桃问题，兔子问题

（1）任务描述

任务一 猴子吃桃问题（40 分）

任务要求：

猴子吃桃问题：猴子第一天摘下若干个桃子，当即吃了一半，还不瘾，又多吃了一个第二天早上又将剩下的桃子吃掉一半，又多吃了一个。以后每天早上都吃了前一天剩下的一半零一个。到第 10 天早上想再吃时，见只剩下一个桃子了，求第一天共摘了多少。

1.正确分析猴子吃桃的规律。

2.程序执行图如下：

图 1-5 猴子吃桃程序执行图

任务二 兔子问题（40 分）

任务要求：

古典问题：有一对兔子，从出生后第 3 个月起每个月都生一对兔子，小兔子长到第三个月后每个月又生一对兔子，假如兔子都不死，输出前 50 个月的兔子总数为多少？提示：（斐波拉契数列）

1.正确分析兔子数量规律

2.提示：（斐波拉契数列），注意算法的选择，防止数据溢出

3.程序执行图如下：

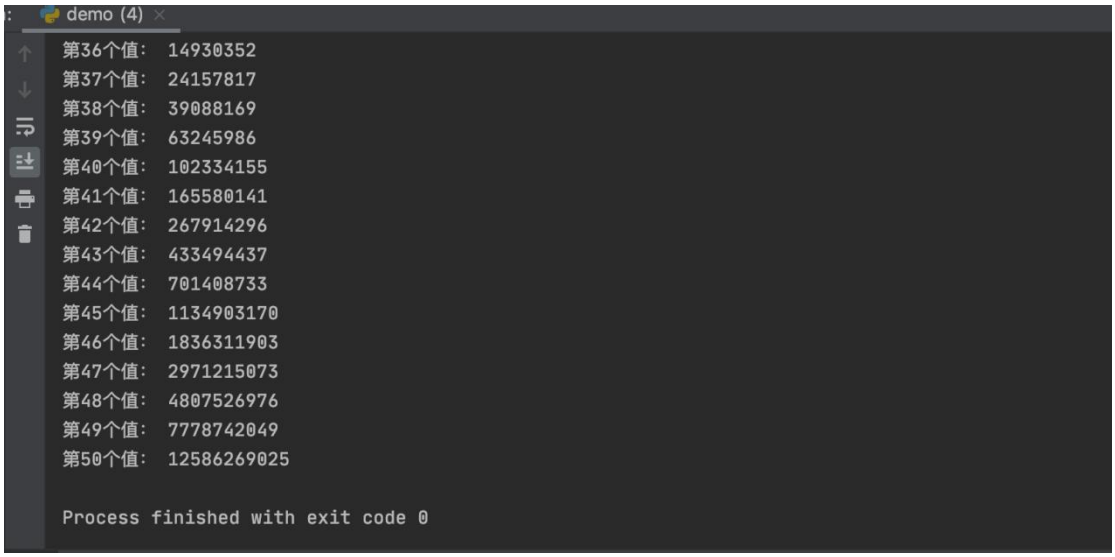


图 1-6 兔子问题执行图（部分月的数据）

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

表 1-5 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	
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120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1-6 考核评价标准

评价内容		分值	评分标准		备注
工作任务一	规律分析	30分	正确使用判定	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。
			正确进行变量声明	10分	
			输出正确桃子数目	10分	
	前后桃子数量分析	5分	桃子数量为（x+1）×2	5分	
	循环语句	5分	判断是否符合要求	5分	
工作任务二	规律分析	30分	正确使用循环判定	12分	
			正确进行变量声明	10分	
			结果输出正确	8分	
	斐波拉契数列	5分	斐波拉契数列高效率实现	5分	
	是否溢出	5分	程序执行效率问题	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：1-1-4，找 1000 以内的所有“水仙花数”，自由落体问题

（1）任务描述

任务一 找 1000 以内的所有“水仙花数”

任务要求：

打印出所有的“水仙花数”，所谓“水仙花数”是指一个三位数，其各位数字立方和等于该数本身。例如：153 是一个“水仙花数”，因为 153=1 的三次方＋5 的三次方＋3 的三次方。任务分析，利用 for 循环控制 100-999 个数，每个数分解出个位，十位，百位。

1.正确分析水仙花数的规律

2.调试程序，程序执行结果如图所示：。

图 1-7 水仙花数计算结果图

任务二 自由落体问题

任务要求：

一球从 100 米高度自由落下，每次落地后反跳回原高度的一半；再落下，求它在第 10 次落地时，共经过多少米？第 10 次反弹多高？任务分析：关键计算，第 n 次落地时共经过的米数，第 n 次反跳高度。

1.任务分析：关键计算，第 n 次落地时共经过的米数，第 n 次反跳高度。

2.寻找规律，如图所示。

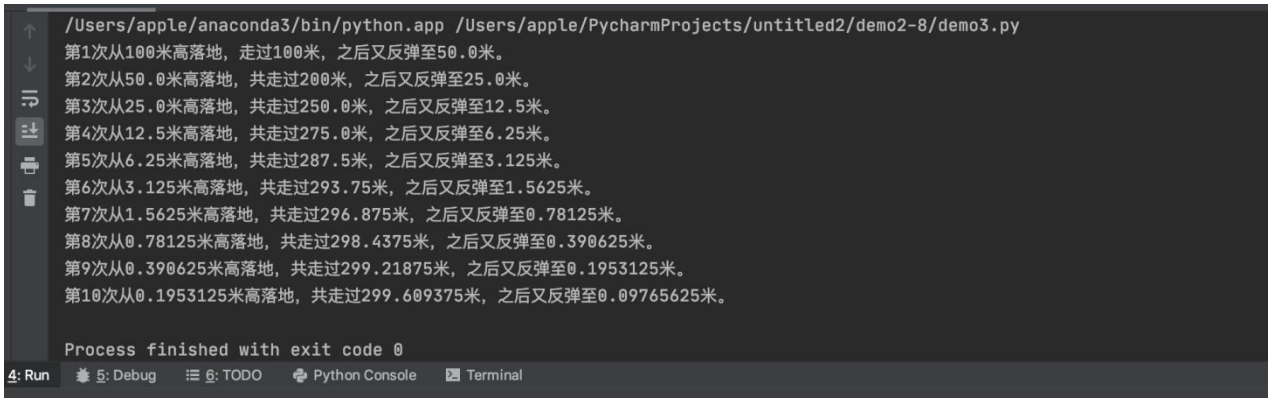


图 1-8　自由落体问题结果

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

表 1-7 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	
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 120 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 1-8 考核评价标准

<table>
<tr><th colspan="2">评价内容</th><th>分值</th><th colspan="2">评分标准</th><th>备注</th></tr>
<tr><td rowspan="5">工作任务一</td><td rowspan="3">规律分析</td><td rowspan="3">30 分</td><td>正确使用循环判定</td><td>10 分</td><td rowspan="12">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。</td></tr>
<tr><td>变量定义正确</td><td>10 分</td></tr>
<tr><td>结果输出正确</td><td>10 分</td></tr>
<tr><td>水仙花树计算</td><td>5 分</td><td>水仙花树范围正确</td><td>5 分</td></tr>
<tr><td>是否溢出</td><td>5 分</td><td>程序执行效率问题</td><td>5 分</td></tr>
<tr><td rowspan="5">工作任务二</td><td rowspan="3">规律分析</td><td rowspan="3">30 分</td><td>正确使用循环判定</td><td>10 分</td></tr>
<tr><td>变量定义正确</td><td>10 分</td></tr>
<tr><td>结果输出正确</td><td>10 分</td></tr>
<tr><td>反弹次数</td><td>5 分</td><td>反弹次数正确</td><td>5 分</td></tr>
<tr><td>是否溢出</td><td>5 分</td><td>程序执行效率问题</td><td>5 分</td></tr>
<tr><td rowspan="2">职业素养</td><td>专业素养</td><td>10 分</td><td>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。</td><td>0-10 分</td></tr>
<tr><td>道德规范</td><td>10 分</td><td>着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场。</td><td>0-10 分</td></tr>
<tr><td colspan="2">总计</td><td colspan="4">100 分</td></tr>
</table>

5.试题编号：1-1-5，求 100 以内所有偶数的和，企业发放奖金问题

（1）任务描述

任务一 求 100 以内所有偶数的和（40 分）

任务要求：

求 1~100 之间所有偶数的和，关键是判断偶数规则，循环并求和。

1.正确分析规律。

2.程序执行结果如下图所示。

图 1-9 1~100 之间所有偶数的和

任务二 企业发放奖金问题（40 分）

任务要求：

企业发放的奖金根据利润提成。利润(I)低于或等于 10 万元时，奖金可提 10%；利润高于 10 万元，低于 20 万元时，低于 10 万元的部分按 10%提成，高于 10 万元的部分，可可提成 7.5%；20 万到 40 万之间时，高于 20 万元的部分，可提成 5%；40 万到 60 万之间时高于 40 万元的部分，可提成 3%；60 万到 100 万之间时，高于 60 万元的部分，可提成 1.5%，高于 100 万元时，超过 100 万元的部分按 1%提成，从键盘输入当月利润的值，求应发放奖金总数，并在控制台输出结果。

1.各阶段奖金计算是否正确。

2.调试程序。

3.程序执行结果如下：

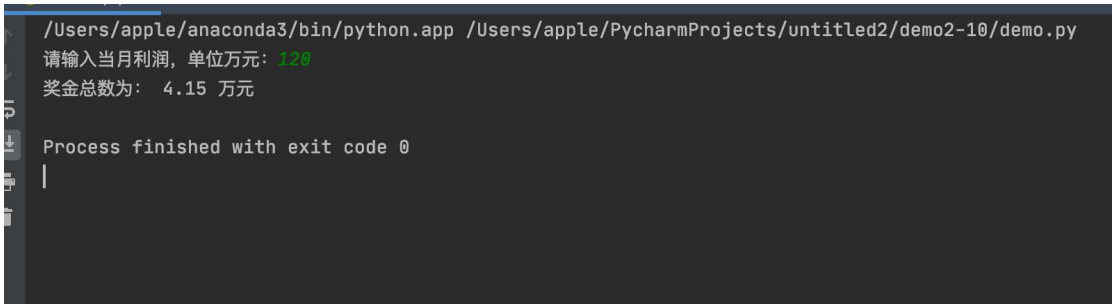


图 1-10 企业发放奖金问题结果图

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

表 1-9 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	
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 120 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 1-10 考核评价标准

评价内容		分值	评分标准		备注
工作任务一	规律分析	30 分	正确使用循环判定	10 分	1、考试舞弊、抄袭、没有按要求
			变量定义正确	10 分	
			结果输出正确	10 分	

	范围分析	5 分	范围大小定义正确	5 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。
	是否溢出	5 分	程序执行效率问题	5 分	
工作任务二	规律分析	30 分	正确使用循环判定	10 分	
			变量定义正确	10 分	
			结果输出正确	10 分	
	范围分析	5 分	范围大小定义正确	5 分	
	是否溢出	5 分	程序执行效率问题	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			

项目 2：大数据平台部署与开发

6.试题编号：1-2-1，Hadoop 平台完全分布式部署；

（1）任务描述

你作为某公司运维工程师，需安装分布式 hadoop 环境。本项目主要完成 hadoop 环境的搭建，本环节需要使用 root 用户完成相关配置。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（20 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（5 分）；
- 根据项目描述，完成 JDK 的环境配置（10 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH;

③ 使用命令验证 JDK 是否安装成功（5 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（70 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（5 分）；
- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（3 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（30 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置;

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置;

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置;

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置;

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置;

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置;

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置;

- 启动 hadoop 个组件（30 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点;

C. 启动 datanode 节点;

D. 使用 jps -l 命令查看进程是否启动成功;

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（20 分）

序号	评分内容	评分点	分值（分）
1	卸载 open JDK	成功卸载 open jdk	10
2	安装包上传	安装文件成功上传	2
2	安装包解压	安装文件成功解压	2
3	配置环境变量	配置 JAVA_HOME	4
4	环境验证	验证 JDK 安装成功	2

评分项二：Hadoop 环境配置（70 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	2
2	安装包解压	安装文件成功解压	3
3	配置环境变量	配置 HADOOP_HOME	3
4	环境验证	验证 Hadoop 安装成功	2
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	10
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	5
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	5
9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5
10	格式化 namenode	格式化成功	10
11	启动 namenode	成功启动 namenode	8
12	启动 datanode	成功启动 datanode	6
13	验证安装	相关进程	6

7.试题编号：1-2-2，Hadoop 常用命令—创建目录；

（1）任务描述

你作为某公司运维工程师,需部署维护 hadoop 环境。本项目主要完成 hadoop 环境的搭建、HDFS 常用命令创建目录。本环节需要使用 root 用户完成相关配置。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（10 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（5 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（3 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（55 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（2 分）；
- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（6 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（25 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（20 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 HDFS 命令—创建目录（25 分）

- 使用 HDFS 相关命令，列出 HDFS 根目录下的文件(5 分)
- 使用 HDFS 相关命令，创建/user/dfstest 目录（10 分）
- 使用 HDFS 相关命令，创建/user/test/example 目录（10 分）
- 使用 HDFS 相关命令，验证/user 目录下的文件（5 分）

将命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《Hadoop 数据分析任务三答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（10 分）

序号	评分内容	评分点	分值（分）
1	卸载 open JDK	成功卸载 open jdk	5
2	安装包上传	安装文件成功上传	1

2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（55 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	2
2	安装包解压	安装文件成功解压	3
3	配置环境变量	配置 HADOOP_HOME	3
4	环境验证	验证 Hadoop 安装成功	2
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	5
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	5
9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5
10	格式化 namenode	格式化成功	5
11	启动 namenode	成功启动 namenode	5
12	启动 datanode	成功启动 datanode	5
13	验证安装	相关进程	5

评分项三：hdfs 命令（25 分）

序号	评分内容	评分点	分值（分）
1	hdfs 命令查看	正确执行 hdfs 命令	5
2	创建 hdfs 目录	正确执行 hdfs 命令	5
3	递归创建 hdfs 子目录	正确执行 hdfs 命令	10
4	hdfs 命令查看文件	正确执行 hdfs 命令	5

8.试题编号：1-2-3，Hadoop 常用命令—上传文件；

（1）任务描述

你作为某公司运维工程师,需部署维护 hadoop 环境。本项目主要完成 hadoop 环境的搭建、HDFS 常用命令。本环节需要使用 root 用户完成相关配置。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（10 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（5 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（3 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（50 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（2 分）；
- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（6 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（25 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（15 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 HDFS 命令—上传文件（30 分）

- 在本地创建文件 a.txt,分别写入“I have a pen, I have an apple”内容；(5 分)
- 使用 HDFS 相关命令，创建/user/dfstest 目录；（5 分）
- 使用 HDFS 相关命令的-copyFromLocal 参数将本地文件 a.txt 上传到 HDFS 目录/user/dfstest 中；（5 分）
- 使用 HDFS 相关命令的-moveFromLocal 参数将本地文件 a.txt 移动到 HDFS 目录并重命名/user/dfstest/b.txt；（5 分）
- 使用 HDFS 相关命令的-put 参数将本地文件 a.txt 上传到 HDFS 目录并重命名/user/dfstest/c.txt；（5 分）

- 使用 HDFS 相关命令，验证文件是否上传成功。（5 分）

将命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《Hadoop 数据分析任务三答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（10 分）

序号	评分内容	评分点	分值（分）
1	卸载 open JDK	成功卸载 open jdk	5
2	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（50 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	2
2	安装包解压	安装文件成功解压	3
3	配置环境变量	配置 HADOOP_HOME	3
4	环境验证	验证 Hadoop 安装成功	2
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	5
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	5
9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5
10	格式化 namenode	格式化成功	5
11	启动 namenode	成功启动 namenode	2
12	启动 datanode	成功启动 datanode	3
13	验证安装	相关进程	5

评分项三：hdfs 命令（30 分）

序号	评分内容	评分点	分值（分）
1	创建本地文件	正确创建文件并添加内容	5
2	创建 hdfs 目录	正确执行 hdfs 命令	5
3	上传本地到 hdfs	正确执行 hdfs 命令	5
4	上传本地到 hdfs	正确执行 hdfs 命令	5
5	Put 命令上传文件	正确执行 hdfs 命令	5
6	hdfs 命令查看文件	正确执行 hdfs 命令	5

9.试题编号：1-2-4，Hadoop 常用命令—下载文件；

（1）任务描述

你作为某公司运维工程师,需部署维护 hadoop 环境。本项目主要完成 hadoop 环境的搭建、HDFS 常用命令。本环节需要使用 root 用户完成相关配置。

本项目主要完成 hadoop 环境的搭建、选择 hdfs 命令上传文件、编写 MapReduce 程序完成数据分析、hdfs 命令查看计算结果。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（10 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（5 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（3 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（50 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（2 分）；

- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（6 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（25 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（15 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 HDFS 命令—下载文件（30 分）

- 在本地创建文件 a.txt,分别写入“I have a pen, I have an apple”内容；(2 分)

- 使用 HDFS 相关命令，创建/user/dfstest 目录；（5 分）
- 使用 HDFS 相关命令的-put 参数将本地文件 a.txt 上传到 HDFS 目录 /user/dfstest/；（5 分）
- 使用 HDFS 相关命令的-moveFromLocal 参数将本地文件 a.txt 移动到 HDFS 目录并重命名/user/dfstest/b.txt；（5 分）
- 使用 HDFS 相关命令的-copyToLocal 参数将/user/dfstest/a.txt 下载到本地当前目录；（5 分）
- 使用 HDFS 相关命令的-get 参数将/user/dfstest/b.txt 下载到本地当前目录；（5 分）
- 本地验证文件是否下载成功。（3 分）

将命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《Hadoop 数据分析任务三答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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		测评专家满足任一条件

	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）。	

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（10 分）

序号	评分内容	评分点	分值（分）
1	卸载 open JDK	成功卸载 open jdk	5
2	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（50 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	2
2	安装包解压	安装文件成功解压	3
3	配置环境变量	配置 HADOOP_HOME	3
4	环境验证	验证 Hadoop 安装成功	2
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	5
8	配置 yarn-site.xml	正确配置相关的配置项：	5

	文件	yarn.nodemanager.aux-services	
9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5
10	格式化 namenode	格式化成功	5
11	启动 namenode	成功启动 namenode	2
12	启动 datanode	成功启动 datanode	3
13	验证安装	相关进程	5

评分项三：hdfs 命令（30 分）

序号	评分内容	评分点	分值（分）
1	创建本地文件	正确创建文件并添加内容	2
2	创建 hdfs 目录	正确执行 hdfs 命令	5
3	Put 命令上传本地到 hdfs	正确执行 hdfs 命令	5
4	moveFromLocal 命令上传本地到 hdfs	正确执行 hdfs 命令	5
5	Get 命令下载文件	正确执行 hdfs 命令	5
6	copyToLocal 命令下载文件	正确执行 hdfs 命令	5
7	查看文件	正确执行 hdfs 命令	3

评价内容		评分标准		备注
工作任务	安装 JDK	JDK 的环境安装与配置	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。
	Hadoop 安装	Hadoop 环境变量配置	10 分	
		Hadoop 配置文件配置	20 分	
		启动 Hadoop	10 分	
	Hadoop 常用命令—下载文件	创建本地文件	2 分	
		创建目录	5 分	
		上传文件	10 分	
		下载文件	10 分	
		验证	3 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：1-2-5，Hadoop 常用命令—删除文件/目录；

（1）任务描述

你作为某公司运维工程师,需部署维护 hadoop 环境。本项目主要完成 hadoop 环境的搭建、HDFS 常用命令。本环节需要使用 root 用户完成相关配置。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（10 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（5 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（3 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（50 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（2 分）；

- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（6 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（25 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（15 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 HDFS 命令—删除文件/目录（30 分）

- 在本地创建文件 a.txt,分别写入“I have a pen, I have an apple”内容(2 分)

- 使用 HDFS 相关命令，创建/user/dfstest/rmdir 目录；（5 分）
- 使用 HDFS 相关命令的-put 参数将本地文件 a.txt 上传到 HDFS 目录 /user/dfstest/；（5 分）
- 使用 HDFS 相关命令查看/user/dfstest/a.txt 文件内容；（5 分）
- 使用 HDFS 相关命令删除/user/dfstest/a.txt 文件；（5 分）
- 使用 HDFS 相关命令删除/user/dfstest/rmdir 目录；（5 分）
- 相关命令验证目录或文件是否删除成功（3 分）

将命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《Hadoop 数据分析任务三答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。	

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（10 分）

序号	评分内容	评分点	分值（分）
1	卸载 open JDK	成功卸载 open jdk	5
2	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（50 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	2
2	安装包解压	安装文件成功解压	3
3	配置环境变量	配置 HADOOP_HOME	3
4	环境验证	验证 Hadoop 安装成功	2
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	5
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	5
9	配置 hadoop-env.sh、yarn-env.sh、	正确配置 JAVA_HOME 环境变量	5

	mapred-env.sh 文件		
10	格式化 namenode	格式化成功	5
11	启动 namenode	成功启动 namenode	2
12	启动 datanode	成功启动 datanode	3
13	验证安装	相关进程	5

评分项三：hdfs 命令（30 分）

序号	评分内容	评分点	分值（分）
1	创建本地文件	正确创建文件并添加内容	2
2	创建 hdfs 目录	正确执行 hdfs 命令	5
3	Put 命令上传本地到 hdfs	正确执行 hdfs 命令	5
4	查看 hdfs 文件内容	正确执行 hdfs 命令	5
5	删除 hdfs 文件	正确执行 hdfs 命令	5
6	删除 hdfs 目录	正确执行 hdfs 命令	5
7	查看文件	正确执行 hdfs 命令	3

11.试题编号：1-2-6，使用 MapReduce 程序计算学生期末考试各科最高分；

（1）任务描述

某知名高中学校为了更好地进行教学改革，现需要对过去 10 年的各年级的期末考试得分进行统计分析，找出各科的最高分数，并以此分析 10 年间教师教学成果和学生学习能力的趋势分析。需要拟搭建一个小型的 hadoop 集群来存放学校的考试数据，并选择 MapReduce 进行离线统计分析。其每条数据格式为：

张三 chinese 73 2015 年下，数据项之间使用空格隔开。

本项目主要完成 hadoop 环境的搭建、选择 hdfs 命令上传文件、编写 MapReduce 程序完成数据分析、hdfs 命令查看计算结果。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（5 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（2 分）；

配置 JAVA_HOME；

配置 PATH；

③ 使用命令验证 JDK 是否安装成功（1 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（30 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（2 分）；
- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（2 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（1 分）；

- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（20 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（5 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 编写 MapperReducer 程序（35 分）

- 编写 Mapper 程序(10 分)

A. 确定数据的输入输出类型（LongWritable, Text,Text, IntWritable）

B. 读取一行数据，转化为字符串，按照需求进行数据处理；

C. 根据输出类型，确定数据写入 Reducer 阶段的 Key 与 Value（Text，

IntWritable）；

D. 调用 context 的 write 方法，将数据推送到给 Reducer 任务；

③编写 Reducer 程序(10 分)：

A. 根据 Mapper 阶段的输出类型确定 Reducer 程序的输入类型以及其结果数据类型(Text, IntWritable,Text,IntWritable)；

B. 编写 reduce 方法，获取 Mapper 阶段输出数据中相同科目的最高分数；

C. 调用 context 的 write 方法，将求得的每科最高分数传递到 driver 端；

- 编写 MapReduce Driver 程序（15 分）：

A. 通过 Configuration 对象来配置连接 HDFS；

B. 通过 Configuration 对象初始化 hadoop Job 对象；

C. 通过 Job 对象来设置任务名称（setJobName）；

D. 通过 Job 对象来设置 Mapper 类的类名（setMapperClass）；

E. 通过 Job 对象来设置 Reducer 类的类名（setReducerClass）；

F. 通过 Job 对象来设置 Driver 类的类名（setJarByClass）；

G. 通过 Job 对象来设置输入数据格式（setInputFormatClass）；

H. 通过 Job 对象来设置输出数据格式（setOutputFormatClass）；

I. 通过 Job 对象来设置 Mapper 阶段的输出 Key 的格式(setMapOutputKeyClass)；

J. 通过 Job 对象来设置 Mapper 阶段的输出 Value 的格式(setMapOutputValueClass)；

K. 通过 Job 对象来设置 Reducer 阶段的输出 Key 的格式(setOutputKeyClass)；

L. 通过 Job 对象来设置 Reducer 阶段的输出 Value 的格式(setOutputValueClass)；

M.)指定文件的输入路径（FileInputFormat.addInputPath）；

N. 指定结果的输出路径（FileOutputFormat.setOutputPath）

O. 通过 Job 对象执行任务（job.waitForCompletion）；

将编写的 Mapper 程序、Reducer 程序、Driver 程序的源代码放到考生文件夹中。

任务四 执行 MapReduce 任务处理数据并验证（20 分）

- 使用 Maven 插件的 clean 命令完成数据包的清理；

- 使用Maven插件的package命令完成MapReduce程序的打包并上传到服务器；
- 使用 hadoop 命令执行 MapReduce 程序(hadoop jar)

A. 指定求最高分数 MapReduce 程序所在的 jar 包名 ；

B. 指定求最高分数 MapReduce 程序 driver 程序的完整类名；

C. 指定源数据的目录；

D. 指定结果数据的存储目录；

- 使用 HDFS 命令查看结果数据；

将 MapReduce 程序的启动命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《Hadoop 数据分析任务四答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。	

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（5 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（30 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 HADOOP_HOME	2
4	环境验证	验证 Hadoop 安装成功	1
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	2
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	3
9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5

10	格式化 namenode	格式化成功	2
11	启动 namenode	成功启动 namenode	1
12	启动 datanode	成功启动 datanode	1
13	验证安装	相关进程	1

评分项三：编写 MapReducer 程序（35 分）

序号	评分内容	评分点	分值（分）
1	编写 Mapper 程序	正确编写 Mapper 程序	10
2	编写 Reducer 程序	正确编写 Reducer 程序	10
3	编写 Driver 程序	正确编写 Driver 程序	15

评分项四：运行 MapReduce 程序（20 分）

序号	评分内容	评分点	分值（分）
1	程序打包	MapReduce 程序打包成功	4
2	程序包上传	程序包成功上传服务器	4
3	提交 MapReduce 程序	成功提交 MapReduce 程序	4
4	监控程序运行	正确查看程序运行过程	4
5	验证数据结果	正确查看数据分析结果	4

评分项五：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

12.试题编号：1-2-7，使用 MapReduce 程序计算学生期末考试各科最低分；

（1）任务描述

某知名高中学校为了更好地进行教学改革，现需要对过去 10 年的各年级的期末考试得分进行统计分析，找出各科的最低分数，并以此分析 10 年间教师教学成果和学生学习能力的趋势分析，特别是分析最低分数存在的区间。需要拟搭建一个小型的 hadoop 集群来存放学校的考试数据，并选择 MapReduce 进行离线统计分析。其每条数据格式为：

李四 English 67 2020 年上 ，数据项之间使用空格隔开。

本项目主要完成 hadoop 环境的搭建、选择 hdfs 命令上传文件、编写 MapReduce 程序完成数据分析、hdfs 命令查看计算结果。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（5 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（2 分）；

C. 配置 JAVA_HOME；

D. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（1 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（30 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（1 分）；
- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（2 分）；

- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（20 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（5 分）

F. 使用命令初始化 hadoop 运行环境（-format）；

G. 启动 namenode 节点；

H. 启动 datanode 节点；

I. 使用 jps -l 命令查看进程是否启动成功；

J. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 编写 MapperReducer 程序（35 分）

- 编写 Mapper 程序(10 分)

A. 确定数据的输入输出类型（LongWritable, Text,Text, IntWritable）

B. 读取一行数据，转化为字符串，按照需求进行数据处理；

C. 根据输出类型，确定数据写入 Reducer 阶段的 Key 与 Value（Text，IntWritable）；

D. 调用 context 的 write 方法，将数据推送到给 Reducer 任务；

③编写 Reducer 程序(10 分)：

A. 根据 Mapper 阶段的输出类型确定 Reducer 程序的输入类型以及其结果数据类型(Text, IntWritable,Text,IntWritable)；

B. 编写 reduce 方法，获取 Mapper 阶段输出数据中相同科目的最低分数；

C. 调用 context 的 write 方法，将求得的每科最低分数传递到 driver 端；

- 编写 MapReduce Driver 程序（15 分）：

A. 通过 Configuration 对象来配置连接 HDFS；

B. 通过 Configuration 对象初始化 hadoop Job 对象；

C. 通过 Job 对象来设置任务名称（setJobName）；

D. 通过 Job 对象来设置 Mapper 类的类名（setMapperClass）；

E. 通过 Job 对象来设置 Reducer 类的类名（setReducerClass）；

F. 通过 Job 对象来设置 Driver 类的类名（setJarByClass）；

G. 通过 Job 对象来设置输入数据格式（setInputFormatClass）；

H. 通过 Job 对象来设置输出数据格式（setOutputFormatClass）；

I. 通过 Job 对象来设置 Mapper 阶段的输出 Key 的格式(setMapOutputKeyClass)；

J. 通过 Job 对象来设置 Mapper 阶段的输出 Value 的格式(setMapOutputValueClass)；

K. 通过 Job 对象来设置 Reducer 阶段的输出 Key 的格式(setOutputKeyClass)；

L. 通过 Job 对象来设置 Reducer 阶段的输出 Value 的格式(setOutputValueClass)；

M. 指定文件的输入路径（FileInputFormat.addInputPath）；

N. 指定结果的输出路径（FileOutputFormat.setOutputPath）

O. 通过 Job 对象执行任务（job.waitForCompletion）；

将编写的 Mapper 程序、Reducer 程序、Driver 程序的源代码放到考生文件夹中。

任务四 执行 MapReduce 任务处理数据并验证（20 分）

- 使用 Maven 插件的 clean 命令完成数据包的清理；
- 使用 Maven 插件的 package 命令完成 MapReduce 程序的打包并上传到服务器；
- 使用 hadoop 命令执行 MapReduce 程序(hadoop jar)

A. 指定求最低分数 MapReduce 程序所在的 jar 包名 ；

B. 指定求最低分数 MapReduce 程序 driver 程序的完整类名；

C. 指定源数据的目录；

D. 指定结果数据的存储目录；

- 使用 HDFS 命令查看结果数据；

将 MapReduce 程序的启动命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《Hadoop 数据分析任务四答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。	

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（5 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（30 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 HADOOP_HOME	2
4	环境验证	验证 Hadoop 安装成功	1
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	2
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	3
9	配置 hadoop-env.sh、yarn-env.sh、	正确配置 JAVA_HOME 环境变量	5

	mapred-env.sh 文件		
10	格式化 namenode	格式化成功	2
11	启动 namenode	成功启动 namenode	1
12	启动 datanode	成功启动 datanode	1
13	验证安装	相关进程	1

评分项三：编写 MapReducer 程序（35 分）

序号	评分内容	评分点	分值（分）
1	编写 Mapper 程序	正确编写 Mapper 程序	10
2	编写 Reducer 程序	正确编写 Reducer 程序	10
3	编写 Driver 程序	正确编写 Driver 程序	15

评分项四：运行 MapReduce 程序（20 分）

序号	评分内容	评分点	分值（分）
1	程序打包	MapReduce 程序打包成功	4
2	程序包上传	程序包成功上传服务器	4
3	提交 MapReduce 程序	成功提交 MapReduce 程序	4
4	监控程序运行	正确查看程序运行过程	4
5	验证数据结果	正确查看数据分析结果	4

评分项五：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

13.试题编号：1-2-8，使用 MapReduce 程序计算学生期末考试各科平均分；

（1）任务描述

某知名高中学校为了更好地进行教学改革，现需要对过去 10 年的各年级的期末考试得分进行统计分析，找出各科的平均，并以此分析 10 年间教师教学成果和学生学习能力的趋势分析。需要拟搭建一个小型的 hadoop 集群来存放学校的考试数据，并选择 MapReduce 进行离线统计分析。其每条数据格式为：

王五 Physics 80 2019 年下 ，数据项之间使用空格隔开。

本项目主要完成 hadoop 环境的搭建、选择 hdfs 命令上传文件、编写 MapReduce 程序完成数据分析、hdfs 命令查看计算结果。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（5 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（2 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（1 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（30 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（1 分）；
- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（2 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；

- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（20 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ， 完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件， 完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件， 完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件， 完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件， 完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（5 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 编写 MapperReducer 程序（35 分）

- 编写 Mapper 程序(10 分)

A. 确定数据的输入输出类型（LongWritable, Text,Text, IntWritable）

B. 读取一行数据，转化为字符串，按照需求进行数据处理；

C. 根据输出类型，确定数据写入 Reducer 阶段的 Key 与 Value（Text，

IntWritable）；

D. 调用 context 的 write 方法，将数据推送到给 Reducer 任务；

③编写 Reducer 程序(10 分)：

A. 根据 Mapper 阶段的输出类型确定 Reducer 程序的输入类型以及其结果数据类型(Text, IntWritable,Text,IntWritable)；

B. 编写 reduce 方法，获取 Mapper 阶段输出数据中相同科目的平均分数；

C. 调用 context 的 write 方法，将求得的每科平均分数传递到 driver 端；

- 编写 MapReduce Driver 程序（15 分）：

通过 Configuration 对象来配置连接 HDFS；

A. 通过 Configuration 对象初始化 hadoop Job 对象；

B. 通过 Job 对象来设置任务名称（setJobName）；

C. 通过 Job 对象来设置 Mapper 类的类名（setMapperClass）；

D. 通过 Job 对象来设置 Reducer 类的类名（setReducerClass）；

E. 通过 Job 对象来设置 Driver 类的类名（setJarByClass）；

F. 通过 Job 对象来设置输入数据格式（setInputFormatClass）；

G. 通过 Job 对象来设置输出数据格式（setOutputFormatClass）；

H. 通过 Job 对象来设置 Mapper 阶段的输出 Key 的格式(setMapOutputKeyClass)；

I. 通过 Job 对象来设置 Mapper 阶段的输出 Value 的格式(setMapOutputValueClass)；

J. 通过 Job 对象来设置 Reducer 阶段的输出 Key 的格式(setOutputKeyClass)；

K. 通过 Job 对象来设置 Reducer 阶段的输出 Value 的格式(setOutputValueClass)；

L. 指定文件的输入路径（FileInputFormat.addInputPath）；

M. 指定结果的输出路径（FileOutputFormat.setOutputPath）

N. 通过 Job 对象执行任务（job.waitForCompletion）；

将编写的 Mapper 程序、Reducer 程序、Driver 程序的源代码放到考生文件夹中。

任务四 执行 MapReduce 任务处理数据并验证（20 分）

- 使用 Maven 插件的 clean 命令完成数据包的清理；

- 使用Maven插件的package命令完成MapReduce程序的打包并上传到服务器；
- 使用 hadoop 命令执行 MapReduce 程序(hadoop jar)

A. 指定求平均分数 MapReduce 程序所在的 jar 包名 ；

B. 指定求平均分数 MapReduce 程序 driver 程序的完整类名；

C. 指定源数据的目录；

D. 指定结果数据的存储目录；

- 使用 HDFS 命令查看结果数据；

将 MapReduce 程序的启动命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《hadoop 数据分析任务四答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。	

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（5 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（30 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 HADOOP_HOME	2
4	环境验证	验证 Hadoop 安装成功	1
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	2
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	3
9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5

10	格式化 namenode	格式化成功	2
11	启动 namenode	成功启动 namenode	1
12	启动 datanode	成功启动 datanode	1
13	验证安装	相关进程	1

评分项三：编写 MapReducer 程序（35 分）

序号	评分内容	评分点	分值（分）
1	编写 Mapper 程序	正确编写 Mapper 程序	10
2	编写 Reducer 程序	正确编写 Reducer 程序	10
3	编写 Driver 程序	正确编写 Driver 程序	15

评分项四：运行 MapReduce 程序（20 分）

序号	评分内容	评分点	分值（分）
1	程序打包	MapReduce 程序打包成功	4
2	程序包上传	程序包成功上传服务器	4
3	提交 MapReduce 程序	成功提交 MapReduce 程序	4
4	监控程序运行	正确查看程序运行过程	4
5	验证数据结果	正确查看数据分析结果	4

评分项五：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

14.试题编号：1-2-9，使用 MapReduce 程序计算学生期末考试各科区间分布情况；

（1）任务描述

某知名高中学校为了更好地进行教学改革，现需要对过去 10 年的各年级的期末考试得分进行统计分析，统计出优秀、良好、好、及格、不及格的数量，通过其分布图来展示教学质量，并以此分析 10 年间教师教学成果和学生学习能力的趋势分析，特别是优秀占比和不及格占比。需要拟搭建一个小型的 hadoop 集群来存放学校的考试数据，并选择 MapReduce 进行离线统计分析。其每条数据格式为：

陈真 Chemistry 93 2021 年上 ，数据项之间使用空格隔开。

本项目主要完成 hadoop 环境的搭建、选择 hdfs 命令上传文件、编写 MapReduce 程序完成数据分析、hdfs 命令查看计算结果。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（5 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（2 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（1 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（30 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（1 分）；
- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（2 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（20 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ，完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件，完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件，完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（5 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 编写 MapperReducer 程序（35 分）

- 编写 Mapper 程序(10 分)

A. 确定数据的输入输出类型（LongWritable, Text,Text, IntWritable）

B. 读取一行数据，转化为字符串，按照需求进行数据处理，根据分数将其转化为优秀、良好、好、及格、不及格；

C. 根据输出类型，确定数据写入 Reducer 阶段的 Key 与 Value（Text，IntWritable）；

D. 调用 context 的 write 方法，将数据推送到给 Reducer 任务；

③编写 Reducer 程序(10 分)：

A. 根据 Mapper 阶段的输出类型确定 Reducer 程序的输入类型以及其结果数据类型(Text, IntWritable,Text,IntWritable)；

B. 编写 reduce 方法，获取 Mapper 阶段输出数据中相同科目的优秀、良好、好、及格、不及格的数量；

C. 调用 context 的 write 方法，将求得的优秀、良好、好、及格、不及格的数量传递到 driver 端；

- 编写 MapReduce Driver 程序（15 分）：

A. 通过 Configuration 对象来配置连接 HDFS；

B. 通过 Configuration 对象初始化 hadoop Job 对象；

C. 通过 Job 对象来设置任务名称（setJobName）；

D. 通过 Job 对象来设置 Mapper 类的类名（setMapperClass）；

E. 通过 Job 对象来设置 Reducer 类的类名（setReducerClass）；

F. 通过 Job 对象来设置 Driver 类的类名（setJarByClass）；

G. 通过 Job 对象来设置输入数据格式（setInputFormatClass）；

H. 通过 Job 对象来设置输出数据格式（setOutputFormatClass）；

I. 通过 Job 对象来设置 Mapper 阶段的输出 Key 的格式(setMapOutputKeyClass)；

J. 通过 Job 对象来设置 Mapper 阶段的输出 Value 的格式(setMapOutputValueClass)；

K. 通过 Job 对象来设置 Reducer 阶段的输出 Key 的格式(setOutputKeyClass)；

L. 通过 Job 对象来设置 Reducer 阶段的输出 Value 的格式(setOutputValueClass)；

M. 指定文件的输入路径（FileInputFormat.addInputPath）；

N. 指定结果的输出路径（FileOutputFormat.setOutputPath）

O. 通过 Job 对象执行任务（job.waitForCompletion）；

将编写的 Mapper 程序、Reducer 程序、Driver 程序的源代码放到考生文件夹中。

任务四 执行 MapReduce 任务处理数据并验证（20 分）

- 使用 Maven 插件的 clean 命令完成数据包的清理；
- 使用 Maven 插件的 package 命令完成 MapReduce 程序的打包并上传到服务器；
- 使用 hadoop 命令执行 MapReduce 程序(hadoop jar)

A. 指定求分数分布的 MapReduce 程序所在的 jar 包名 ；

B. 指定求分数分布的 MapReduce 程序 driver 程序的完整类名；

C. 指定源数据的目录；

D. 指定结果数据的存储目录；

- 使用 HDFS 命令查看结果数据；

将 MapReduce 程序的启动命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《hadoop 数据分析任务四答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（5 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（30 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 HADOOP_HOME	2
4	环境验证	验证 Hadoop 安装成功	1
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5

7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	2
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	3
9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5
10	格式化 namenode	格式化成功	2
11	启动 namenode	成功启动 namenode	1
12	启动 datanode	成功启动 datanode	1
13	验证安装	相关进程	1

评分项三：编写 MapReducer 程序（35 分）

序号	评分内容	评分点	分值（分）
1	编写 Mapper 程序	正确编写 Mapper 程序	10
2	编写 Reducer 程序	正确编写 Reducer 程序	10
3	编写 Driver 程序	正确编写 Driver 程序	15

评分项四：运行 MapReduce 程序（20 分）

序号	评分内容	评分点	分值（分）
1	程序打包	MapReduce 程序打包成功	4
2	程序包上传	程序包成功上传服务器	4
3	提交 MapReduce 程序	成功提交 MapReduce 程序	4
4	监控程序运行	正确查看程序运行过程	4
5	验证数据结果	正确查看数据分析结果	4

评分项五：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

15.试题编号：1-2-10，使用 MapReduce 程序计算学生期末考试各科成绩进行排序；

（1）任务描述

某知名高中学校为了更好地进行教学改革，现需要对过去 10 年的各年级的期末考试得分进行统计分析，过滤掉成绩为空的数据并对各科成绩进行排序，并以此分析 10 年间教师教学成果和学生学习能力的趋势分析，特别是分析最低分数存在的区间。需要拟搭建一个小型的 hadoop 集群来存放学校的考试数据，并选择 MapReduce 进行离线统计分析。其每条数据格式为：

李明 English 67 2011 年上 ，数据项之间使用空格隔开。

本项目主要完成 hadoop 环境的搭建、选择 hdfs 命令上传文件、编写 MapReduce 程序完成数据分析、hdfs 命令查看计算结果。

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+Hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 Hadoop 数据分析 01 张三。

任务一 搭建 JDK 环境（5 分）

- 根据项目描述，完成 JDK 安装文件的上传与解压（2 分）；
- 根据项目描述，完成 JDK 的环境配置（1 分）；

A. 配置 JAVA_HOME；

B. 配置 PATH；

③ 使用命令验证 JDK 是否安装成功（1 分）；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务一答案.doc》,文件内容格式如下：

JDK 安装文件的解压命令是： XXXXXX，并给出截图；

JDK 的环境配置内容是：xxxxx ，给出截图；

JDK 安装正确验证命令是：xxxxx 给出截图；

将该答案文件保存到考生文件夹中。

任务二 搭建 Hadoop 环境（30 分）

- 根据任务要求，将 Hadoop 的安装文件上传到服务器并解压（1 分）；

- 根据任务要求，配置 hadoop 的环境变量，包含 bin 和 sbin（2 分）；
- 使用命令验证 HADOOP 环境变量是否配置成功（2 分）；
- 进入 hadoop 目录下的 etc/hadoop 子目录中进行文件的配置（20 分）；

A. 配置 core-site.xml 文件，完成 fs.defaultFS、hadoop.tmp.dir 配置项的配置；

B. 配置 hdfs-site.xml 文件，完成 dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address 配置项的配置；

C. 配置 mapred-site.xml 文件 ， 完成 mapreduce.framework.name 配置项的配置；

D. 配置 yarn-site.xml 文件， 完成 yarn.nodemanager.aux-services 配置项的配置；

E. 配置 hadoop-env.sh 文件， 完成 JAVA_HOME 环境变量的配置；

F. 配置 yarn-env.sh 文件， 完成 JAVA_HOME 环境变量的配置；

G. 配置 mapred-env.sh 文件， 完成 JAVA_HOME 环境变量的配置；

- 启动 hadoop 个组件（5 分）

A. 使用命令初始化 hadoop 运行环境（-format）；

B. 启动 namenode 节点；

C. 启动 datanode 节点；

D. 使用 jps -l 命令查看进程是否启动成功；

E. 打开参数 dfs.namenode.http-address 配置页面查看是否可以打开；

将该任务的答案存放到答案文件中，文件命名为《Hadoop 数据分析任务二答案.doc》,文件内容格式如下：

Hadoop 环境变量内容是：xxxxx,给出截图；

Hadoop 环境验证命令是：xxxxxx，给出截图

Hadoop 的 namenode、datanode 启动命令是：xxxxx，给出运行成功截图；

Hadoop 的 namenode 管理界面的截图；

任务三 编写 MapperReducer 程序（35 分）

- 编写 Mapper 程序(10 分)

A. 确定数据的输入输出类型（LongWritable, Text,Text, IntWritable）

B. 读取一行数据，转化为字符串，按照需求进行数据处理，将分数项为空的过滤掉；

C. 根据输出类型，确定数据写入 Reducer 阶段的 Key 与 Value（Text，IntWritable）；

D. 调用 context 的 write 方法，将数据推送到给 Reducer 任务；

③编写 Reducer 程序(10 分)：

A. 根据 Mapper 阶段的输出类型确定 Reducer 程序的输入类型以及其结果数据类型(Text, IntWritable,Text,IntWritable)；

B. 编写 reduce 方法，获取 Mapper 阶段输出数据中相同科目的分数并按照从大到小进行排序；

C. 调用 context 的 write 方法，将求得的每科排序后的数据递到 driver 端；

- 编写 MapReduce Driver 程序（15 分）：

A. 通过 Configuration 对象来配置连接 HDFS；

B. 通过 Configuration 对象初始化 hadoop Job 对象；

C. 通过 Job 对象来设置任务名称（setJobName）；

D. 通过 Job 对象来设置 Mapper 类的类名（setMapperClass）；

E. 通过 Job 对象来设置 Reducer 类的类名（setReducerClass）；

F. 通过 Job 对象来设置 Driver 类的类名（setJarByClass）；

G. 通过 Job 对象来设置输入数据格式（setInputFormatClass）；

H. 通过 Job 对象来设置输出数据格式（setOutputFormatClass）；

I. 通过 Job 对象来设置 Mapper 阶段的输出 Key 的格式(setMapOutputKeyClass)；

J. 通过 Job 对象来设置 Mapper 阶段的输出 Value 的格式(setMapOutputValueClass)；

K. 通过 Job 对象来设置 Reducer 阶段的输出 Key 的格式(setOutputKeyClass)；

L. 通过 Job 对象来设置 Reducer 阶段的输出 Value 的格式(setOutputValueClass)；

M. 指定文件的输入路径（FileInputFormat.addInputPath）；

N. 指定结果的输出路径（FileOutputFormat.setOutputPath）

O. 通过 Job 对象执行任务（job.waitForCompletion）；

将编写的 Mapper 程序、Reducer 程序、Driver 程序的源代码放到考生文件夹中。

任务四 执行 MapReduce 任务处理数据并验证（20 分）

- 使用 Maven 插件的 clean 命令完成数据包的清理；
- 使用 Maven 插件的 package 命令完成 MapReduce 程序的打包并上传到服务器；
- 使用 hadoop 命令执行 MapReduce 程序(hadoop jar)

A）指定分数排序 MapReduce 程序所在的 jar 包名 ；

B）指定分数排序 MapReduce 程序 driver 程序的完整类名；

C) 指定源数据的目录；

D）指定结果数据的存储目录；

- 使用 HDFS 命令查看结果数据；

将 MapReduce 程序的启动命令和运行界面截图以及查看结果数据的命令和执行结果存放到答案文件中，答案文件命名为《hadoop 数据分析任务四答案.doc》，并将答案文件存放到考生文件夹中。

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+hadoop 数据分析+考生号+考生姓名，示例：湖南信息职业技术学院 hadoop 数据分析 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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	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		测评专家满足

专家	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）。	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 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：JDK 环境搭建（5 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 JAVA_HOME	2
4	环境验证	验证 JDK 安装成功	1

评分项二：Hadoop 环境配置（30 分）

序号	评分内容	评分点	分值（分）
1	安装包上传	安装文件成功上传	1
2	安装包解压	安装文件成功解压	1
3	配置环境变量	配置 HADOOP_HOME	2
4	环境验证	验证 Hadoop 安装成功	1
5	配置 core-site.xml 文件	正确配置相关的配置项：df.defaultFS、hadoop.tmp.dir	5
6	配置 hdfs-site.xml 文件	正确配置相关的配置项：dfs.replication、dfs.namenode.name.dir、dfs.datanode.data.dir、dfs.namenode.http-address	5
7	配置 mapred-site.xml 文件	正确配置相关的配置项：mapreduce.framework.name	2
8	配置 yarn-site.xml 文件	正确配置相关的配置项：yarn.nodemanager.aux-services	3

9	配置 hadoop-env.sh、yarn-env.sh、mapred-env.sh 文件	正确配置 JAVA_HOME 环境变量	5
10	格式化 namenode	格式化成功	2
11	启动 namenode	成功启动 namenode	1
12	启动 datanode	成功启动 datanode	1
13	验证安装	相关进程	1

评分项三：编写 MapReducer 程序（35 分）

序号	评分内容	评分点	分值（分）
1	编写 Mapper 程序	正确编写 Mapper 程序	10
2	编写 Reducer 程序	正确编写 Reducer 程序	10
3	编写 Driver 程序	正确编写 Driver 程序	15

评分项四：运行 MapReduce 程序（20 分）

序号	评分内容	评分点	分值（分）
1	程序打包	MapReduce 程序打包成功	4
2	程序包上传	程序包成功上传服务器	4
3	提交 MapReduce 程序	成功提交 MapReduce 程序	4
4	监控程序运行	正确查看程序运行过程	4
5	验证数据结果	正确查看数据分析结果	4

评分项五：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

二、岗位核心技能

模块二 数据采集

项目 1：基于 Flume 的数据采集

16.试题编号：2-1-1，使用 Flume 采集指定文件数据

（1）任务描述

某公司随着业务的扩展，其运营研发部提出开发新产品来支撑运营，其主要是通过收集不同产品的运行日记，也支持工程师将指定的文件数据纳入到采集系统中。先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需要支持工程师手动或者使用命令来将文件数据发送到 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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件

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（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（2 分）；

将该任务的答案存放到答案文件中，文件命名为《数据采集任务一答案.doc》，文件内容格式如下：

Flume source 组件为： xxx

Flume channel 组件为： xxx

Flume sink 组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①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 hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent 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

①根据任务描述和拓扑图,创建 agent 的配置文件,并取名为 agent.properties,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。(3 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。（2 分）

③编写 avro source 组件配置项(10 分)：

A）配置 source 组件的类型标识配置项(type)；

B）配置 source 组件监听的 IP(bind)；

C）配置 source 组件监听的端口(port)；

④编写 memory channel 组件的配置项（10 分）：

A）配置 channel 组件的类型标识配置项(type)；

B）配置 channel 组件容量大小配置项(capacity)；

C）配置 channel 组件事务容量大小配置项(transactionCapacity)；

⑤编写 logger sink 组件的配置项（10 分）：

A）配置 sink 组件的类型标识配置项（type）；

B）配置 sink 组件的最大显示信息长度(maxBytesToLog)

⑥将创建好的 flume 组件组装为完整的 agent（10 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

B) 配置 sink 组件需要连接的 channel 的配置项（channel）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用 Flume 命令启动 agent 处理数据

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 Flume agent，并且通过 -Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台。(10 分)

②使用 Flume 的安装目录下 bin 目录下的 flume-ng 脚本，通过 avro-client 参

数指定启用 avro 客户端来发送数据，指定需要发送的文件（-F）,文件发送到哪台服务器(-H)，发送到指定服务器的哪个端口(-p) （5 分）。

将 agent 的启动命令和运行界面截图以及发送指定文件的命令和执行结果存放到答案文件中，答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中。

任务五 验证数据是否正确处理

①logger 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，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，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

<table>
<tr><th>项目</th><th colspan="2">基本实施条件</th><th>备注</th></tr>
<tr><td>场地</td><td colspan="2">能同时容纳 30 人以上现场考核</td><td></td></tr>
<tr><td rowspan="2">设备</td><td colspan="2">30 台以上的主流计算机
安装 Centos7 或更高版本</td><td>用于程序设计，每人一台。</td></tr>
<tr><td colspan="2">FTP 服务器 1 台</td><td>用于保存测试人员考试结果</td></tr>
<tr><td>工具</td><td>开发工具</td><td>XShell、SecureCRT</td><td>用以连接服务器</td></tr>
<tr><td rowspan="2">测评专家</td><td colspan="2">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）。</td><td rowspan="2">测评专家满足任一条件</td></tr>
<tr><td colspan="2">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）。</td></tr>
</table>

（3）考核时量

考核时间为 120 分钟

（4）评分细则

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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	3
3	sink 类型选择	选择正确的 sink 组件	2

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	2
2	组件别名定义	声明各个组件唯一别名	3

评分项四：Flume source 配置（10 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	4
2	source bind 配置	配置正确的 bind 值	3
3	source port 配置	配置正确的 port 值	3

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	3
2	channel capacity 配置	配置正确的容量值	3
3	channel 事务容量配置	配置正确的事务容量值	4

评分项六：Flume sink 配置（10 分）

序号	评分内容	评分点	分值（分）
1	sink type 配置	配置正确的 type 值	5
2	sink maxBytesToLog 配置	配置正确的显示长度值	5

评分项七：Flume agent 连接配置（10 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	5
2	sink 连接 channel	sink 正确连接 channel	5

评分项八：Flume agent 启动与验证（25 分）

序号	评分内容	评分点	分值（分）
1	agent 启动命令	agent 启动成功	10
2	flume-ng 发送数据	agent 成功发送数据	5
3	source 采集数据处理验证	agent source 成功采集数据	5
4	sink 取出数据验证	agent sink 成功取出数据	5

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

17.试题编号：2-1-2，使用 Flume 采集 shell 命令或者脚本的结果数据

（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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件

①根据项目描述，选择能够支持处理指定文件的 Flumesource 组件（5 分）；

②根据项目描述，选择能够进行快速测试的 Flumechannel 组件（3 分）；

③根据项目描述，选择能够实时展示测试数据的 Flumesink 组件（2 分）；

将该任务的答案存放到答案文件中，文件命名为《数据采集任务一答案.doc》，文件内容格式如下：

Flume source 组件为： xxx

Flume channel 组件为： xxx

Flume sink 组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①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 hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent 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

①根据任务描述和拓扑图,创建 agent 的配置文件,并取名为 agent.properties,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。(3 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。（2 分）

③编写 exec source 组件配置项(10 分)：

A）配置 source 组件的类型标识配置项(type)；

B）配置 source 组件监听的 Unix 命令或者脚本(command)；

④编写 Memory channel 组件的配置项（10 分）：

A）配置 channel 组件的类型标识配置项(type)；

B）配置 channel 组件容量大小配置项(capacity)；

C）配置 channel 组件事务容量大小配置项(transactionCapacity)；

⑤编写 logger sink 组件的配置项（10 分）：

A）配置 sink 组件的类型标识配置项(type)；

B）配置 sink 组件的最大显示信息长度(maxBytesToLog)

⑥将创建好的 flume 组件组装为完整的 agent（10 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

B) 配置 sink 组件需要连接的 channel 的配置项（channel）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用 Flume 命令启动 agent 处理数据

①使用 Flume 的安装目录下 bin 目录下的 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 Flume agent，并且通过-Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台（10 分）。

将 agent 的启动命令和运行界面截图以及发送指定文件的命令和执行结果存放到答案文件中，答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中（5 分）。

任务五 验证数据是否正确处理

①logger 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，答案文件命名为《数据采集任务五答

案.doc》，并将答案文件存放到考生文件夹中(10 分)。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-2-1 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 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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	3
3	sink 类型选择	选择正确的 sink 组件	2

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	2
2	组件别名定义	声明各个组件唯一别名	3

评分项四：Flume source 配置（10 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	5
2	source command 配置	配置正确的 command 值	5

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	3
2	channel capacity 配置	配置正确的容量值	3
3	channel 事务容量配置	配置正确的事务容量值	4

评分项六：Flume sink 配置（10 分）

序号	评分内容	评分点	分值（分）
1	sink type 配置	配置正确的 type 值	5
2	sink maxBytesToLog 配置	配置正确的显示长度值	5

评分项七：Flume agent 连接配置（10 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	5
2	sink 连接 channel	sink 正确连接 channel	5

评分项八：Flume agent 启动与验证（25 分）

序号	评分内容	评分点	分值（分）
1	agent 启动命令	agent 启动成功	10
2	command 命令执行	command 命令执行成功	5
3	source 采集数据处理验证	agent source 成功采集数据	5
4	sink 取出数据验证	agent sink 成功取出数据	5

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

18.试题编号：2-1-3，使用 Flume 采集通过 TCP 协议发送的数据

（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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件

①根据项目描述，选择能够支持处理 TCP 协议发送的数据的 Flume source 组件（5 分）；

②根据项目描述，选择有高可靠性 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（2 分）；

将该任务的答案存放到答案文件中,文件命名为《数据采集任务一答案.doc》,文件内容格式如下：

Flume source 组件为： xxx

Flume channel 组件为： xxx

Flume sink 组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①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 hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent 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

①根据任务描述和拓扑图，创建 agent 的配置文件，取名为 agent.properties,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。(3 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。（2 分）

③编写 netcat tcp source 组件配置项(10 分)：

A）配置 source 组件的类型标识配置项(type)；

B）配置 source 组件的监听的 IP 配置项(bind)；

C）配置 source 组件监听的端口配置项(port)；

④编写 file channel 组件的配置项（10 分）：

A）配置 channel 组件的类型标识配置项(type)；

B）配置 channel 组件数据缓存目录配置项(dataDirs)；

C）配置 channel 组件元数据 checkpoint 缓存目录配置项(checkpointDir)；

D）配置 channel 组件容量大小配置项(capacity)；

E）配置 channel 组件事务容量大小配置项(transactionCapacity)；

⑤编写 logger sink 组件的配置项（10 分）：

A）配置 sink 组件的类型标识配置项(type)；

B）配置 sink 组件的最大显示信息长度(maxBytesToLog)

⑥将创建好的 flume 组件组装为完整的 agent（10 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

B) 配置 sink 组件需要连接的 channel 的配置项（channel）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用 Flume 命令启动 agent 处理数据

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 Flume agent，并且通过 -Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台。

②使用 telnet 命令使用 TCP 协议发送日报、周报结构化数据(10 分)。

将 agent 的启动命令和运行界面截图以及 telnet 命令发送数据成功的截图和执行结果存放到答案文件中，答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中（5 分）。

任务五 验证数据是否正确处理

①loggersink 会将读取到数据打印到控制台，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，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-3-1 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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	3
3	sink 类型选择	选择正确的 sink 组件	2

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	2
2	组件别名定义	声明各个组件唯一别名	3

评分项四：Flume source 配置（10 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	3
2	source bind 配置	配置正确的 ip 值	4
3	source port 配置	配置正确的 port 值	3

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	2
2	channel capacity 配置	配置正确的容量值	2
3	channel 事务容量配置	配置正确的事务容量值	2
4	channel dataDirs 配置	配置正确的存储路径值	2
5	channel checkpointDir 配置	配置正确的检查点路径值	2

评分项六：Flume sink 配置（10 分）

序号	评分内容	评分点	分值（分）
1	sink type 配置	配置正确的 type 值	5
2	sink maxBytesToLog 配置	配置正确的显示长度值	5

评分项七：Flume agent 连接配置（10 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	5
2	sink 连接 channel	sink 正确连接 channel	5

评分项八：Flume agent 启动与验证（25 分）

序号	评分内容	评分点	分值（分）
1	agent 启动命令	agent 启动成功	10
2	Telnet 命令执行	Telnet 命令成功发送数据	5
3	source 采集数据处理验证	agent source 成功采集数据	5
4	sink 取出数据验证	agent sink 成功取出数据	5

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

19.试题编号：2-1-4，使用 Flume 采集通过运营支撑数据

（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。同时为了支持优化其自动运维系统，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，需要收集这些自动运维脚本或者命令的运行结果数据。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以及接收 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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件

①根据项目描述，选择能够支持处理 TCP 协议发送的数据以及 shell 脚本的结果数据的 Flume source 组件（5 分）；

②根据项目描述，选择有高可靠性的 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（2 分）；

将该任务的答案存放到答案文件中，文件命名为《数据采集任务一答案.doc》，文件内容格式如下：

Flume source 组件为： xxx、xxx

Flume channel 组件为： xxx

Flume sink 组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①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 hniu。将画出的拓扑图截图并命名为“数据采集

任务二：agent 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

①根据任务描述和拓扑图,创建 agent 的配置文件,并取名为 agent.properties,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。(3 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。（2 分）

③编写 Netcat Tcp source 组件配置项(10 分)：

A）配置 source 组件的类型标识配置项（type）；

B）配置 source 组件的监听的 IP 配置项（bind）；

C）配置 source 组件监听的端口配置项（port）；

④编写 exec source 组件配置项(5 分)：

A）配置 source 组件的类型标识配置项（type）；

B）配置 source 组件监听的 Unix 命令或者脚本（command）；

⑤编写 File channel 组件的配置项（10 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件数据缓存目录配置项（dataDirs）；

C）配置 channel 组件元数据 checkpoint 缓存目录配置项(checkpointDir)；

D）配置 channel 组件容量大小配置项(capacity)；

E）配置 channel 组件事务容量大小配置项(transactionCapacity)；

⑥编写 Logger sink 组件的配置项（5 分）：

A）配置 sink 组件的类型标识配置项(type)；

B）配置 sink 组件的最大显示信息长度(maxBytesToLog)

⑦将创建好的 flume 组件组装为完整的 agent（5 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

B) 配置 sink 组件需要连接的 channel 的配置项（channel）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用 Flume 命令启动 agent 处理数据

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 Flume agent，并且通过 -Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台（10 分）。

②使用 telnet 命令使用 TCP 协议发送日报、周报结构化数据（5 分）。

将 agent 的启动命令和运行界面截图以及 telnet 命令发送数据成功的截图和执行结果存放到答案文件中，答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中（5 分）。

任务五 验证数据是否正确处理

①loggersink 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，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中（5 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-4-1 项目实施条件

<table>
<tr><th>项目</th><th colspan="2">基本实施条件</th><th>备注</th></tr>
<tr><td>场地</td><td colspan="2">能同时容纳 30 人以上现场考核</td><td></td></tr>
<tr><td rowspan="2">设备</td><td colspan="2">30 台以上的主流计算机
安装 Centos7 或更高版本</td><td>用于程序设计，每人一台。</td></tr>
<tr><td colspan="2">FTP 服务器 1 台</td><td>用于保存测试人员考试结果</td></tr>
<tr><td>工具</td><td>开发工具</td><td>XShell、SecureCRT</td><td>用以连接服务器</td></tr>
<tr><td rowspan="2">测评专家</td><td colspan="2">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）。</td><td rowspan="2">测评专家满足任一条件</td></tr>
<tr><td colspan="2">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</td></tr>
</table>

	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 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择所有正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	3
3	sink 类型选择	选择正确的 sink 组件	2

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（10 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	5
2	组件别名定义	声明各个组件唯一别名	5

评分项四：Flume source 配置（15 分）

序号	评分内容	评分点	分值（分）
1	source1 type 配置	配置正确的 type 值	3
2	source1 bind 配置	配置正确的 ip 值	4
3	source1 port 配置	配置正确的 port 值	3
4	source2 type 配置	配置正确的 type 值	3
5	source2 command 配置	配置正确的 command 命令	2

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	2
2	channel capacity 配置	配置正确的容量值	2
3	channel 事务容量配置	配置正确的事务容量值	2
4	channel dataDirs 配置	配置正确的存储路径值	2
5	channel checkpointDir 配置	配置正确的检查点路径值	2

评分项六：Flume sink 配置（5 分）

序号	评分内容	评分点	分值（分）
1	sink type 配置	配置正确的 type 值	3
2	sink maxBytesToLog 配置	配置正确的显示长度值	2

评分项七：Flume agent 连接配置（5 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	3
2	sink 连接 channel	sink 正确连接 channel	2

评分项八：Flume agent 启动与验证（25 分）

序号	评分内容	评分点	分值（分）
1	agent 启动命令	agent 启动成功	10
2	Telnet 命令执行	Telnet 命令成功发送数据	5
3	source 采集数据处理验证	agent source 成功采集数据	5
4	sink 取出数据验证	agent sink 成功取出数据	5

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

20.试题编号：2-1-5，日记数据采集到HDFS的数据采集系统

（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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集01张三。

任务一 选择合适的Flume组件

①根据项目描述，选择能够处理指定目录下文件的Flume source组件（4分）；

②根据项目描述，选择能够缓存数据到磁盘的Flume channel组件（3分）

③根据项目描述，选择能够将数据写入HDFS的Flume sink组件（3分）；

将该任务的答案存放到答案文件中,文件命名为《数据采集任务一答案.doc》,文件内容格式如下：

Flume source组件为： xxx

Flume channel组件为： xxx

Flume sink组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出agent的拓扑图

①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hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent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

①根据任务描述和拓扑图,创建 agent 的配置文件,并取名为 agent.properties,确定所使用的 Flume source 组件的别名、Flume channel 组件的别名。(3 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。（2 分）

③编写 spooling director ysource 组件配置项(15 分)：

A）配置 source 组件的类型标识配置项（type）；

B）配置 source 组件监听的目录配置项（spoolDir）；

C）配置 source 组件处理完成文件后的后缀名配置项（fileSuffix）；

D）配置 source 组件匹配出需要的处理文件名正则表达式配置项（includePattern）；

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）;

F)配置 source 组件将处理的文件名写入到 Event 头部的配置项（basenameHeader）；

G）配置 source 组件将文件名写入头部的哪个字段的配置项（basenameHeaderKey）；

H）配置 source 组件递归扫描监听的文件目录（recursiveDirectorySearch）

④编写 File channel 组件的配置项（10 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件数据缓存目录配置项（dataDirs）；

C）配置 channel 组件元数据 checkpoint 缓存目录配置项(checkpointDir)；

D）配置 channel 组件容量大小配置项(capacity)；

E）配置 channel 组件事务容量大小配置项(transactionCapacity)；

⑤编写 HDFS sink 组件的配置项（15 分）：

A）配置 sink 组件的类型标识配置项(type)；

B）配置 sink 组件的数据写入的 HDFSURL 配置项(hdfs.path)；

C）配置 sink 组件将数据写入 HDFS 完成后文件后缀配置项(hdfs.fileSuffix)；

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)；

E）配置 sink 组件每批次处理数据量的批次大小（hdfs.batchSize）；

F）配置 sink 组件使用本地时间（hdfs.useLocalTimeStamp）；

⑥将创建好的 Flume 组件组装为完整的 agent（5 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

B) 配置 sink 组件需要连接的 channel 的配置项（channel）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用 Flume 命令启动 agent 处理数据

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 Flume agent，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和执行结果存放到答案文件中，答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

任务五 验证数据是否正确处理

①使用 hdfs 命令查看文件是否成功写入到 hdfs，并使用命令查看写入的内容是否正确。将查看文件写入的命令以及查看正确写入数据的截图存入到答案文件中，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中。（10 分）

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-5-1 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	用于程序设计，

	安装 Centos7 或更高版本		每人一台。
	FTP 服务器 1 台		用于保存测试人员考试结果
	Kafka 集群（5 台 broker）		用以创建 topic
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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择所有正确的 source 组件	4
2	channel 类型选择	选择正确的 channel 组件	3
3	sink 类型选择	选择正确的 sink 组件	3

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	3
2	组件别名定义	声明各个组件唯一别名	2

评分项四：Flume source 配置（15 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	2
2	source spoolDir 配置	配置正确的 spoolDir 值	2
3	source fileSuffix 配置	配置正确的 fileSuffix 值	1
4	source includePattern 配置	配置正确 includePattern 值	3
5	source ignorePattern 配置	配置正确的 ignorePattern 命令	3
6	source basenameHeader 配置	配置正确的 basenameHeader 值	1
7	source basenameHeaderKey 配置	配置正确的 basenameHeaderKey 值	1
8	source recursiveDirectorySearch 配置	配置正确的 recursiveDirectorySearch 命令	2

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	2
2	channel capacity 配置	配置正确的容量值	2
3	channel 事务容量配置	配置正确的事务容量值	2
4	channel dataDirs 配置	配置正确的存储路径值	2
5	channel checkpointDir 配置	配置正确的检查点路径值	2

评分项六：Flume sink 配置（15 分）

序号	评分内容	评分点	分值（分）
1	sink type 配置	配置正确的 type 值	3
2	sink hdfs.path 配置	配置正确的 hdfs 路径	3
3	sink hdfs.fileSuffix 配置	配置正确的文件后缀	2
4	sink hdfs.rollInterval 配置	配置正确的间隔时间	2
5	sink hdfs.batchSize 配置	配置正确的批大小	2
6	sink hdfs.useLocalTimeStampe 配置	配置正确的使用本地时间	3

评分项七：Flume agent 连接配置（5 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	3
2	sink 连接 channel	sink 正确连接 channel	2

评分项八：Flume agent 启动与验证（20 分）

序号	评分内容	评分点	分值（分）
1	agent 启动命令	agent 启动成功	10
2	source 采集数据处理验证	agent source 成功采集数据	5
3	sink 取出数据验证	agent sink 成功取出数据并上传至 hdfs 中	5

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

项目 2：基于 kafka 的消息队列数据采集

21.试题编号：2-2-1，单 source、单 channel 构建数据采集系统

（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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件--不使用 Sink

①根据项目描述，选择能够处理指定目录下文件的 Flume source 组件（5 分）；

②根据项目描述，选择能够创建 Kafka Topic 的 Flume channel 组件（5 分）；

将该任务的答案存放到答案文件中，文件命名为《数据采集任务一答案.doc》，文件内容格式如下：

Flume source 组件为： xxx

Flume channel 组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①根据任务一选取的 Flume source 和 channel 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent 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

①根据任务描述和拓扑图，创建 agent 的配置文件，取名为 agent.properties，确定所使用的 Flume source 组件的别名、Flume channel 组件的别名。(3 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。（2 分）

③编写 spooling directory source 组件配置项(15 分)：

A）配置 source 组件的类型标识配置项（type）；

B）配置 source 组件监听的目录配置项（spoolDir）；

C）配置 source 组件处理完成文件后的后缀名配置项（fileSuffix）；

D）配置 source 组件匹配出需要的处理文件名正则表达式配置项（includePattern）；

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）；

F）配置 source 组件递归扫描监听的文件目录（recursiveDirectorySearch）

④编写 Kafka channel 组件的配置项（10 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件服务器 IP 和端口配置项（kafka.bootstrap.servers）；

C）配置 channel 组件数据写入的 Topic 名称配置项（kafka.topic）；

D）配置 channel 组件解析数据类型配置项（parseAsFlumeEvent）

⑤将创建好的 flume 组件组装为完整的 agent（5 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用命令根据 agent 的配置文件，创建 KafkaTopic

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，参数（--zookeeper），用来指定需要连接的 zookeeper 地址。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）（10 分）。

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，答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中。

任务五 使用 Flume 命令启动 agent 处理数据

①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 Flume agent，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，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中（15 分）。

任务六 验证数据是否正确处理

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，答案文件命名为《数据采集任务六答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-6-1 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
	Kafka 集群（5 台 broker）		用以创建 topic
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）。	

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择所有正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	5

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	3
2	组件别名定义	声明各个组件唯一别名	2

评分项四：Flume source 配置（15 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	2
2	source spoolDir 配置	配置正确的 spoolDir 值	2
3	source fileSuffix 配置	配置正确的 fileSuffix 值	1
4	source includePattern 配置	配置正确 includePattern 值	3
5	source ignorePattern 配置	配置正确的 ignorePattern 命令	3
6	source basenameHeader 配置	配置正确的 basenameHeader 值	1
7	source basenameHeaderKey 配置	配置正确的 basenameHeaderKey 值	1

8	source recursiveDirectorySearch 配置	配置正确的 recursiveDirectorySearch 命令	2

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	3
2	channel 连接 Kafka 配置	配置正确的 kafka 服务器地址	3
3	channel 数据存储配置	配置正确的存储 topic	2
4	channel 数据解析配置	配置正确的解析参数值	2

评分项六：Flume agent 连接配置（5 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	5

评分项七：Flume agent 启动与验证（35 分）

序号	评分内容	评分点	分值（分）
1	创建 Kafka topic	Topic 成功创建	10
2	agent 启动命令	agent 启动成功	10
3	source 采集数据处理验证	agent source 成功采集数据	5
4	Kafka 消费数据	Kafka consumer 成功消费数据	10

评分项八：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

22.试题编号：2-2-2，单 source、多 channel 构建数据采集系统

（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、行人结构化数据并以文件的形式进行保存，格式为 txt，如车辆数据文件名为 vehicle.txt，行人数据文件名为 pedestrian.txt。需要把生成的新文件数据实时分别采集到消息队列不同的 topic 中去。

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、创建若干个 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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件--不使用 Sink

①根据项目描述，选择能够处理指定目录下文件的 Flume source 组件（5 分）；

②根据项目描述，选择能够创建 Kafka Topic 的 Flume channel 组件（5 分）；

将该任务的答案存放到答案文件中，文件命名为《数据采集任务一答案.doc》，文件内容格式如下：

Flume source 组件为： xxx

Flume channel 组件为： xxx、xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①根据任务一选取的 Flume source 和 channel 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent 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

①根据任务描述和拓扑图，创建 agent 的配置文件，并取名为 agent.properties，确定所使用的 Flume source 组件的别名、Flume channel 组件的别名。(3 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。（2 分）

③编写 spooling directory source 组件配置项(15 分)：

A）配置 source 组件的类型标识配置项（type）；

B）配置 source 组件监听的目录配置项（spoolDir）；

C）配置 source 组件处理完成文件后的后缀名配置项（fileSuffix）；

D）配置 source 组件匹配出需要的处理文件名正则表达式配置项（includePattern）；

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）；

F)配置 source 组件将处理的文件名写入到 Event 头部的配置项（basenameHeader）；

G）配置 source 组件将文件名写入头部的哪个字段的配置项（basenameHeaderKey）；

H）配置 source 组件递归扫描监听的文件目录（recursiveDirectorySearch）

④编写 Kafkachannel（1）组件的配置项（5 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件服务器 IP 和端口配置项（kafka.bootstrap.servers）；

C）配置 channel 组件数据写入的 Topic 名称配置项（kafka.topic）；

⑤编写 Kafkachannel（2）组件的配置项（5 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件服务器 IP 和端口配置项（kafka.bootstrap.servers）；

C）配置 channel 组件数据写入的 Topic 名称配置项（kafka.topic）；

⑥编写类型为分发 channel selector（multiplexing）相关的配置（5 分）：

A）配置 channel selector 的类型标识配置项（type）；

B）配置 channel selector 基于 header 分发的字段 key（header）；

C）配置 channel selector 将车辆数据分发的 channel（mapping.vehicle.txt）；

D）配置 channel selector 将行人数据分发的 channel（mapping.pedestrian.txt）；

⑦将创建好的 flume 组件组装为完整的 agent（5 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用命令根据 agent 的配置文件，创建 Kafka Topic

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，参数（--zookeeper），用来指定需要连接的 zookeeper 地址。创建一个备份数(--replication-factor) 为 1，分区数（--partitions）为 1 的 topic 存放车辆数据，也需要创建一个备份数为 1，分区数为 1 的 topic 存放行人的数据。这两个 topic 的名字为 agent 配置文件中 channel（1）模块、channel(2)模块中配置的 topic 名称，通过参数（--topic）来指定（10 分）。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,答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中。

任务五 使用 Flume 命令启动 agent 处理数据

①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，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 Flume agent，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，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

任务六 验证数据是否正确处理

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）验证车辆数据和行人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，答案文件命名为《数据采集任务六答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学

院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-7-1 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
	Kafka 集群（5 台 broker）		用以创建 topic
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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择所有正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	5

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	3
2	组件别名定义	声明各个组件唯一别名	2

评分项四：Flume source 配置（15 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	2
2	source spoolDir 配置	配置正确的 spoolDir 值	2
3	source fileSuffix 配置	配置正确的 fileSuffix 值	1
4	source includePattern 配置	配置正确 includePattern 值	3
5	source ignorePattern 配置	配置正确的 ignorePattern 命令	3
6	source basenameHeader 配置	配置正确的 basenameHeader 值	1
7	source basenameHeaderKey 配置	配置正确的 basenameHeaderKey 值	1
8	source recursiveDirectorySearch 配置	配置正确的 recursiveDirectorySearch 命令	2

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel1 type 配置	配置正确的 type 值	2
2	channel1 连接 Kafka 配置	配置正确的 kafka 服务器地址	2
3	channel1 数据存储配置	配置正确的存储 topic	1
4	channel1 type 配置	配置正确的 type 值	2
5	channel2 连接 Kafka 配置	配置正确的 kafka 服务器地址	2
6	channel2 数据存储配置	配置正确的存储 topic	1

评分项六：Flume channel selector 配置（5 分）

序号	评分内容	评分点	分值（分）
1	selector 类型配置	配置正确的 type 值	2
2	selector 分发字段配置	配置正确的字段值	1
3	selector 分发条件一配置	配置正确的分发条件	1

4	selector 分发条件二配置	配置正确的分发条件	1

评分项七：Flume agent 连接配置（5 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	5

评分项八：Flume agent 启动与验证（30 分）

序号	评分内容	评分点	分值（分）
1	创建 Kafka topic	Topic 成功创建	10
2	agent 启动命令	agent 启动成功	10
3	Kafka 消费数据	Kafka consumer 成功消费数据	10

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

23.试题编号：2-2-3，多 source、多 channel 构建数据采集系统

（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，如车辆数据文件名为 vehicle.txt，在完成视频处理分析完后，通过 socket 发送出当前视频的元数据，后续的分析程序即可完成多维度数据关联分析。通过构建多 source、多 channel 把这些生成的数据实时采集到消息队列中去。

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、创建若干个 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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件--不使用 Sink

①根据项目描述，选择能够处理指定目录下文件的 Flume source 组件（5 分）；

②根据项目描述，选择能够接收通过 TCP 协议发送数据的 Flume source 组件（2 分）；

③根据项目描述，选择能够创建 KafkaTopic 的 Flume channel 组件（3 分）；

将该任务的答案存放到答案文件中，文件命名为《数据采集任务一答案.doc》，文件内容格式如下：

Flume source 组件为： xxx、xxx

Flume channel 组件为： xxx、xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①根据任务一选取的 Flume source 和 channel 组件，画出相应的 agent 的拓扑图，并将 agent 命名为 hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent 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

①根据任务描述和拓扑图，创建 agent 的配置文件，取名为 agent.properties，确定所使用的 Flume source 组件的别名、Flume channel 组件的别名。(3 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。（2 分）

③编写 netcat TCP source 组件配置项（5 分）：

A）配置 source 组件的类型标识配置项(type)；

B）配置 source 组件的监听的 IP 配置项(bind)；

C）配置 source 组件监听的端口配置项(port)；

④编写 spooling directory source 组件配置项(15 分)：

A）配置 source 组件的类型标识配置项（type）；

B）配置 source 组件监听的目录配置项（spoolDir）；

C）配置 source 组件处理完成文件后的后缀名配置项（fileSuffix）；

D）配置 source 组件匹配出需要的处理文件名正则表达式配置项（includePattern）；

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）；

F)配置 source 组件将处理的文件名写入到 Event 头部的配置项（basenameHeader）；

G）配置 source 组件将文件名写入头部的哪个字段的配置项（basenameHeaderKey）；

H）配置 source 组件递归扫描监听的文件目录（recursiveDirectorySearch）

⑤编写 kafkachannel（1）组件的配置项（5 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件服务器 IP 和端口配置项（kafka.bootstrap.servers）；

C）配置 channel 组件数据写入的 Topic 名称配置项（kafka.topic）；

⑥编写 kafkachannel（2）组件的配置项（5 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件服务器 IP 和端口配置项（kafka.bootstrap.servers）；

C）配置 channel 组件数据写入的 Topic 名称配置项（kafka.topic）；

⑦将创建好的 flume 组件组装为完整的 agent（5 分）

A) 配置 netcat tcp source 的 channel 组件配置项（channels）

B) 配置 spooling directory source 的 channel 组件配置项（channels）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用命令根据 agent 的配置文件，创建 Kafka Topic

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，参数（--zookeeper），用来指定需要连接的 zookeeper 地址。创建一个备份数(--replication-factor) 为 1，分区数（--partitions）为 1 的 topic 存放车辆数据，也需要创建一个备份数为 1，分区数为 1 的 topic 存放行人的数据。这两个 topic 的名字为 agent 配置文件中 channel（1）模块、channel(2)模块中配置的 topic 名称，通过参数（--topic）来指定(10 分)。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,答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中。

任务五 使用 Flume 命令启动 agent 处理数据

①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，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（10 分）。

②使用 telnet 命令使用 TCP 协议发送视频元数据。

将 agent 的启动命令以及 telnet 命令发送数据成功的截图和运行界面截图执行结果存放到答案文件中，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中（5 分）。

任务六 验证数据是否正确处理

①使用 Kafka 的 bin 目录下的 kafka-console-consumer.sh 命令，通过参数(--bootstrap-server)指定连接的 Kafka 服务器 IP 和端口、需要读取的 Topic(--topic)以及从什么位置开始消费（--from-beginning）验证车辆数据和视频元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，答案文件命名为《数据采集任务六答案.doc》，并将答

案文件存放到考生文件夹中（5 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-8-1 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
	Kafka 集群（5 台 broker）		用以创建 topic
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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择所有正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	5

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	3
2	组件别名定义	声明各个组件唯一别名	2

评分项四：Flume source 配置（20 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	2
2	source spoolDir 配置	配置正确的 spoolDir 值	2
3	source fileSuffix 配置	配置正确的 fileSuffix 值	1
4	source includePattern 配置	配置正确 includePattern 值	3
5	source ignorePattern 配置	配置正确的 ignorePattern 命令	3
6	source basenameHeader 配置	配置正确的 basenameHeader 值	1
7	source basenameHeaderKey 配置	配置正确的 basenameHeaderKey 值	1
8	source recursiveDirectorySearch 配置	配置正确的 recursiveDirectorySearch 命令	2
9	source2 的类型配置	配置正确的 type 值	2
10	source2 的 ip 配置	配置正确的 ip 值	2
11	source2 的 port 配置	配置正确的 port 值	1

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	3
2	channel 连接 Kafka 配置	配置正确的 kafka 服务器地址	4
3	channel 数据存储配置	配置正确的存储 topic	3

评分项六：Flume agent 连接配置（5 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	5

评分项七：Flume agent 启动与验证（30 分）

序号	评分内容	评分点	分值（分）
1	创建 Kafka topic	Topic 成功创建	10
2	agent 启动命令	agent 启动成功	10
3	Telnet 命令发送数据	Telenet 成功发送数据	5
4	Kafka 消费数据	Kafka consumer 成功消费数据	5

评分项八：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

24.试题编号：2-2-4，多 source、单 channel 构建数据采集系统

（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，并且在完成视频处理分析完后，通过 socket 发送当前视频的元数据，通过构建多 source、单 channel 把这些数据实时采集到消息队列中去，并且将收集到的消息以日记的形式实时展示在控制台。

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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

任务一 选择合适的 Flume 组件

①根据项目描述，选择能够处理指定目录下文件的 Flume source 组件（3 分）；

②根据项目描述，选择能够接收通过 TCP 协议发送数据的 Flume source 组件（2 分）；

③根据项目描述，选择能够创建 KafkaTopic 的 Flume channel 组件（3 分）；

④根据项目描述，选择能够把数据以日记形式打印到控制台的 Flumesink 组件（2 分）；

将该任务的答案存放到答案文件中，文件命名为《数据采集任务一答案.doc》，文件内容格式如下：

Flume source 组件为： xxx、xxx

Flume channel 组件为： xxx

Flume sink 组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出 agent 的拓扑图

①根据任务一选取的 Flume source 和 channel 组件、sink 组件，画出相应的 agent

的拓扑图，并将 agent 命名为 hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent 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

①根据任务描述和拓扑图，创建 agent 的配置文件，取名为 agent.properties，确定所使用的 Flum esource 组件的别名、Flume channel 组件的别名以及 Flume sink 组件的别名。(3 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。（2 分）

③编写 netcat TCP source 组件配置项（5 分）：

A）配置 source 组件的类型标识配置项(type)；

B）配置 source 组件的监听的 IP 配置项(bind)；

C）配置 source 组件监听的端口配置项(port)；

④编写 spooling directory source 组件配置项(15 分)：

A）配置 source 组件的类型标识配置项（type）；

B）配置 source 组件监听的目录配置项（spoolDir）；

C）配置 source 组件处理完成文件后的后缀名配置项（fileSuffix）；

D）配置 source 组件匹配出需要的处理文件名正则表达式配置项（includePattern）；

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）；

F)配置 source 组件将处理的文件名写入到 Event 头部的配置项（basenameHeader）；

G）配置 source 组件将文件名写入头部的哪个字段的配置项（basenameHeaderKey）；

H）配置 source 组件递归扫描监听的文件目录（recursiveDirectorySearch）

⑤编写 kafka channel 组件的配置项（5 分）：

A）配置 channel 组件的类型标识配置项（type）；

B）配置 channel 组件服务器 IP 和端口配置项（kafka.bootstrap.servers）；

C）配置 channel 组件数据写入的 Topic 名称配置项（kafka.topic）；

⑥编写 logger sink 组件的配置项（5 分）：

A）配置 sink 组件的类型标识配置项(type)；

B）配置 sink 组件的最大显示信息长度(maxBytesToLog)

⑦将创建好的 flume 组件组装为完整的 agent（5 分）

A) 配置 source 组件需要连接的 channel 的配置项（channels）

B) 配置 sink 组件需要连接的 channel 的配置项（channel）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用命令根据 agent 的配置文件，创建 KafkaTopic

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，参数（--zookeeper），用来指定需要连接的 zookeeper 地址。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,答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

任务五 使用 Flume 命令启动 agent 处理数据

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 Flume agent，并且通过 -Dflume.root.logger=INFO,console 来把 agent 的运行日记打印到控制台（10 分）。

②使用 telnet 命令使用 TCP 协议发送视频元数据（5 分）。

将 agent 的启动命令以及 telnet 命令发送数据成功的截图和运行界面截图执行结果存放到答案文件中，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中。

任务六 验证数据是否正确处理

①使用 Kafka 的 bin 目录下的 kafka-console-consumer.sh 命令，通过参数(--bootstrap-server)指定连接的 Kafka 服务器 IP 和端口、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，答案文件命名为《数据采集任务六答案.doc》，并将答案文件存放到考生文件夹中（5 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-9-1 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
	Kafka 集群（5 台 broker）		用以创建 topic
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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成

恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择所有正确的 source 组件	5
2	channel 类型选择	选择正确的 channel 组件	5

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	3
2	组件别名定义	声明各个组件唯一别名	2

评分项四：Flume source 配置（20 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	2
2	source spoolDir 配置	配置正确的 spoolDir 值	2
3	source fileSuffix 配置	配置正确的 fileSuffix 值	1
4	source includePattern 配置	配置正确 includePattern 值	3
5	source ignorePattern 配置	配置正确的 ignorePattern 命令	3
6	source basenameHeader 配置	配置正确的 basenameHeader 值	1
7	source basenameHeaderKey 配置	配置正确的 basenameHeaderKey 值	1
8	source recursiveDirectorySearch 配置	配置正确的 recursiveDirectorySearch 命令	2
9	source2 的类型配置	配置正确的 type 值	2
10	source2 的 ip 配置	配置正确的 ip 值	2
11	source2 的 port 配置	配置正确的 port 值	1

评分项五：Flume channel 配置（5 分）

序号	评分内容	评分点	分值（分）
1	channel1 type 配置	配置正确的 type 值	2
2	channel1 连接 Kafka 配置	配置正确的 kafka 服务器地址	2

3	channel1 数据存储配置	配置正确的存储 topic	1

评分项六：Flume sink 配置（5 分）

序号	评分内容	评分点	分值（分）
1	sink type 配置	配置正确的 type 值	3
2	sink maxBytesToLog 配置	配置正确的显示长度值	2

评分项七：Flume agent 连接配置（5 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	3
2	sink 连接 channel	sink 正确连接 channel	2

评分项八：Flume agent 启动与验证（30 分）

序号	评分内容	评分点	分值（分）
1	创建 Kafka topic	Topic 成功创建	10
2	agent 启动命令	agent 启动成功	10
3	Telnet 命令发送数据	Telenet 成功发送数据	5
4	Kafka 消费数据	Kafka consumer 成功消费数据	5

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

25.试题编号：2-2-5，日记数据采集到消息队列的采集系统

（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。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集01张三。

任务一 选择合适的Flume组件

①根据项目描述，选择能够处理指定目录下文件的Flume source组件（4分）；

②根据项目描述，选择能够缓存数据到磁盘的Flume channel组件（3分）

③根据项目描述，选择能够创建Kafka Topic的 Flume sink组件（3分）；

将该任务的答案存放到答案文件中,文件命名为《数据采集任务一答案.doc》,文件内容格式如下：

Flume source组件为： xxx

Flume channel组件为： xxx

Flume sink组件为： xxx

将该答案文件保存到考生文件夹中。

任务二 画出agent的拓扑图

①根据任务一选取的Flumes ource和channel、sink组件，画出相应的agent的拓扑图，并将agent命名为hniu。将画出的拓扑图截图并命名为“数据采集任务二：agent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

①根据任务描述和拓扑图,创建agent的配置文件,并取名为agent.properties,

确定所使用的Flume source 、Flume channel 、Flume sink组件的别名。(3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。（2分）

③编写spooling directory source组件配置项(15分)：

A）配置source组件的类型标识配置项（type）；

B）配置source组件监听的目录配置项（spoolDir）；

C）配置source组件处理完成文件后的后缀名配置项（fileSuffix）；

D）配置source组件匹配出需要的处理文件名正则表达式配置项（includePattern）；

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）；

F)配置source组件将处理的文件名写入到Event头部的配置项（basenameHeader）；

G）配置source组件将文件名写入头部的哪个字段的配置项（basenameHeaderKey）；

H）配置source组件递归扫描监听的文件目录（recursiveDirectorySearch）

④编写File channel组件的配置项（10分）：

A）配置channel组件的类型标识配置项（type）；

B）配置channel组件数据缓存目录配置项（dataDirs）；

C）配置channel组件元数据checkpoint缓存目录配置项(checkpointDir)；

D）配置channel组件容量大小配置项(capacity)；

E）配置channel组件事务容量大小配置项(transactionCapacity)；

⑤编写Kafka sink组件的配置项（10分）：

A）配置sink组件的类型标识配置项（type）；

B）配置sink组件的连接Kafka集群的IP和端口(kafka.bootstrap.servers)；

C）配置sink组件将数据读出写入的Topic配置项(kafka.topic)

⑥将创建好的Flume组件组装为完整的agent（5分）

A) 配置source组件需要连接的channel的配置项（channels）

B) 配置sink组件需要连接的channel的配置项（channel）

将编写完成的配置文件保存到考生文件夹里。

任务四 使用命令根据 agent 的配置文件，创建 KafkaTopic

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，参数（--zookeeper），用来指定需要连接的 zookeeper 地址。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）(10 分)。

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，答案文件命名为《数据采集任务四答案.doc》，并将答案文件存放到考生文件夹中。

任务五 使用 Flume 命令启动 agent 处理数据

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 Flume agent，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，答案文件命名为《数据采集任务五答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

任务六 验证数据是否正确处理

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，答案文件命名为《数据采集任务六答案.doc》，并将答案文件存放到考生文件夹中（10 分）。

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+数据采集+考生号+考生姓名，示例：湖南信息职业技术学院数据采集 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，代码源文件以“姓名_题号”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

表 2-10-1 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
	Kafka 集群（5 台 broker）		用以创建 topic
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）。		

（3）考核时量

考核时间为 120 分钟

（4）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 10%，工作任务完成质量占该项目总分的 90%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：Flume 组件选型（10 分）

序号	评分内容	评分点	分值（分）
1	source 类型选择	选择所有正确的 source 组件	4
2	channel 类型选择	选择正确的 channel 组件	3
3	sink 类型选择	选择正确的 sink 组件	3

评分项二：Flume agent 拓扑图（10 分）

序号	评分内容	评分点	分值（分）
1	拓扑图结构	画出正确的拓扑图	5
2	组件类型标识	拓扑图各组件标识正确	5

评分项三：Flume 组件声明（5 分）

序号	评分内容	评分点	分值（分）
1	配置文件创建	使用命令成果创建配置文件	3
2	组件别名定义	声明各个组件唯一别名	2

评分项四：Flume source 配置（15 分）

序号	评分内容	评分点	分值（分）
1	source type 配置	配置正确的 type 值	2
2	source spoolDir 配置	配置正确的 spoolDir 值	2
3	source fileSuffix 配置	配置正确的 fileSuffix 值	1
4	source includePattern 配置	配置正确 includePattern 值	3
5	source ignorePattern 配置	配置正确的 ignorePattern 命令	3
6	source basenameHeader 配置	配置正确的 basenameHeader 值	1
7	source basenameHeaderKey 配置	配置正确的 basenameHeaderKey 值	1
8	source recursiveDirectorySearch 配置	配置正确的 recursiveDirectorySearch 命令	2

评分项五：Flume channel 配置（10 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	2
2	channel capacity 配置	配置正确的容量值	2
3	channel 事务容量配置	配置正确的事务容量值	2
4	channel dataDirs 配置	配置正确的存储路径值	2
5	channel checkpointDir 配置	配置正确的检查点路径值	2

评分项六：Flume sink 配置（5 分）

序号	评分内容	评分点	分值（分）
1	channel type 配置	配置正确的 type 值	2
2	channel 连接 Kafka 配置	配置正确的 kafka 服务器地址	2

3	channel 数据存储配置	配置正确的存储 topic	1

评分项七：Flume agent 连接配置（5 分）

序号	评分内容	评分点	分值（分）
1	source 连接 channel	source 正确连接 channel	3
2	sink 连接 channel	sink 正确连接 channel	2

评分项八：Flume agent 启动与验证（30 分）

序号	评分内容	评分点	分值（分）
1	创建 Kafka topic	Topic 成功创建	10
2	agent 启动命令	agent 启动成功	10
3	Telnet 命令发送数据	Telenet 成功发送数据	5
4	Kafka 消费数据	Kafka consumer 成功消费数据	5

评分项九：职业素质（10 分）

序号	评分内容	评分点	分值（分）
1	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	5
2	道德素养	着装干净、整洁。举止文明，遵守考场纪律，按顺序进出考场	5

模块三 数据清洗与挖掘应用

项目 1：基于 kettle 的数据清洗

26.试题编号：3-1-1：Excel 数据清洗

（1）任务描述

某学校计算机学院分别采集课程开课记录数据（01_1_inputdata.xlsx）和学生基本信息数据（01_3_inputdata.xls）各字段说明如下表 1 和表 3 所示，

课程开课记录源数据有数据缺失,不准确等问题，要求利用 Kettle 软件对该数据先进行转换，然后进行利用参照表数据进行课程的类别校验，最后将查询过的数据导出，课程名称与课程类别的参照表数据（01_2_inputdata_Ref.xlsx）如下表 2 所示。

学生基本信息源数据由于数据录入出错、数据不完整等原因，会导致集成后同一实体对应多条记录，在数据清洗的过程中，重复记录的检测与清除是一项十分重要的工作。要求利用 Kettle 软件对该数据进行相应的清洗并且导出。

表 1 课程开课记录数据字段说明

字段名称	位置	说明	示例
ID	1	编号	1
Name	2	课程名称	数据预处理技术
Category	3	课程类别	基础课
Class_hours	4	课时数	48

表 2 课程参照表数据字段说明

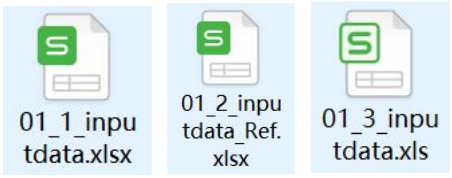
课程名称	课程类别
大学英语	基础课
数据预处理技术	专业课
Python 基础编程	专业课
Hadoop 大数据基础	专业课

表 3 学生基本信息数据字段说明

字段名称	位置	说明	示例
姓名	1	唱片标题	张珊

性别	2	艺术家	女
出生日期	3	发行国家	19990601
班级	4	唱片公司	大数据 1901
QQ 号码	5	唱片价钱	102452221

源数据文档名称：01_1_inputdata.xlsx，01_2_inputdata_Ref.xlsx，01_3_inputdata.xls。



以下任务需要按照步骤对配置参数界面进行截图，所有截图保存到物理机上指定位置“E:\技能抽查提交资料\考生学校+考生号+考生姓名\T01 答案.docx”，具体操作见提交要求。

任务 1 课程开课记录数据的校验

任务 1.1：转换文件的新建与数据的导入（5 分）

1. 在 Kettle 软件中新建名为 01_1 的转换文件（.ktr），选择输入类别中的 Excel 输入步骤.

2. 配置 Excel 输入步骤中的向相关参数，将源数据 01_1_inputdata.xlsx 导入 Kettle 软件中。

3. 再次选择 Excel 输入步骤，配置 Excel 输入 2 步骤中的向相关参数，将源数据 01_2_inputdata_Ref.xlsx 导入 Kettle 软件中。

任务 1.2：数据的清洗（30 分）

1. 选择查询类别中的流查询步骤，建立 Excel 输入步骤和 Excel 输入 2 步骤到流查询步骤之间的连接。

2. 设置流查询步骤中 lookup step：Excel 输入 2，查询值所需的关键字：字段为‘Name’，查询字段为‘课程名称’，指定用来接收的字段：Field 为‘课程类别’，新的名称为‘课程 REF’，默认为‘******’，类型为‘String’。

3. 选择转换类别中的计算器步骤，建立流查询步骤到计算器步骤之间的连接。设置计算器步骤参数，新字段为‘计算分数’，计算为‘Jaro similitude between Sting A and Sting B’,字段 A 为‘Categroy’,字段 B 为‘课程 REF’,

值类型为‘None’,移除为‘否’。

任务 1.3：数据的导出和维护（5 分）

1.选择输出类别中的 Microsoft Excel 输出步骤，建立计算器步骤到 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，将查询过的数据以名为 01_1_outputdata.xlsx 数据文件类型导出。

任务 2 学生基本信息数据的去重

任务 2.1：数据库准备，转换文件的新建与数据的导入（5 分）

1. 在 Kettle 软件中新建名为 01_2 的转换文件（.ktr），选择输入类别中的 Excel 输入步骤.

2. 配置 Excel 输入步骤中的向相关参数，将源数据 01_3_inputdata.xls 导入 Kettle 软件中。

任务 2.2：数据的清洗（30 分）

1. 基于姓名字段进行模糊匹配出有可能相似的数据。复制一份 Excel 输入步骤,得到 Excel 输入 2，择查询类别中的模糊匹配步骤，建立 Excel 输入步骤和 Excel 输入 2 步骤到模糊匹配步骤之间的连接。设置模糊匹配步骤相关参数，实现匹配出不完全重复记录。在一般选项中，匹配步骤设置为表输入 2，匹配字段设为姓名，主要流字段设为姓名，算法设置为 Jaro，最小值为 0，最大值为 0.8。在字段选项中指定额外的在匹配流中的字段，选择编号字段改名为匹配编号，选择 QQ 号码改名为匹配 QQ。

2. 根据 QQ 号码字段过滤出不完全重复数据。选择流程类别中的过滤记录步骤，建立模糊匹配步骤到过滤记录步骤之间的连接。设置过滤记录步骤条件为：‘QQ 号码’=‘匹配 QQ’，发送 true 数据给步骤：过滤记录 2。

3. 对过滤得到的不完全重复数据进行保留一条数据的操作。再次选择流程类别中的过滤记录步骤，建立过滤记录步骤到过滤记录 2 步骤之间的连接，要求 Result is TRUE。设置过滤记录 2 步骤条件为：‘编号’<‘匹配编号’。

任务 2.3：数据的导出和维护（5 分）

1. 选择输出类别中的 Microsoft Excel 输出步骤，建立过滤记录步骤到 Microsoft Excel 输出步骤之间的连接，要求 Result is FLASE。建立过滤记录

2 步骤到 Microsoft Excel 输出步骤之间的连接，要求 Result is TRUE。

2. 配置 Microsoft Excel 步骤中相关参数，将清洗过的数据以字段为姓名，性别，出生日期，班级，QQ 号码的顺序导出名为 02_2_outputdata.xls 的数据文件类型。

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：湖南信息职业技术学院 01 张三，并且在考生文件夹中创建“T01 答案.docx”文件。

2）“技能抽查提交资料”文件夹内保存截图 word 文档、转换源文件及引用的相关数据文件，转换源文件以“姓名_题号.ktr”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

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 8GB 以上,硬盘 100G	无

②软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7	安装 64 位版本
2	Kettle	12.0 及以上	导入 mysql 的 jar 包
3	Mysql	5.7	1. 开放进行外部连接权限 2. 时区设置为“+8:00”
4	Navicat for MySQL	11.0 及以上	无
5	Microsoft Office	2007 及以上	无

（3）考核时量

考核时间为 150 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 4：

表 4 数据清洗模块考核评价标准

评价内容			配分	评分标准		备注
工作任务	模块一	数据的导入	5 分	数据输入步骤的选择符合要求	2	1、考试舞弊、抄袭、

				两个源数据按照要求导入kettle中	3	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。
		数据的清洗	30 分	流查询步骤的选择和连接符合要求	2	
				流查询步骤参数设置符合要求	14	
				计算器步骤的选择和连接符合要求。	2	
				计算器步骤的设置符合要求	10	
		数据的导出和维护	5 分	Microsoft Excel 输出步骤选择符合要求	2	
				数据正确导出为 xlsx 文件	3	
	模块二	数据库准备和数据的导入	5 分	数据输入步骤的选择符合要求	2	
				源数据按照要求导入 kettle 中	3	
		数据的清洗	30 分	正确复制一份表输入步骤，模糊匹配步骤选择符合要求	2	
				正确实现匹配出不完全重复记录	10	
				两次过滤记录步骤的选择符合要求，过滤记录步骤到过滤记录2步骤之间的连接符合要求	2	
				过滤记录步骤条件设置符合要求	8	
				过滤记录2步骤条件设置符合要求	8	
		数据的导出和维护	5 分	Microsoft Excel 输出步骤的选择和连接符合要求	2	
				数据按照要求正确导出为xls文件	5	
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			

27.试题编号：3-1-2：TXT 数据和 XML 数据清洗

（1）任务描述

某唱片外企采集的职工基本信息数据（02_1_inputdata.txt）和唱片信息数据（02_2_inputdata.xml）各字段说明如下表 1 和表 2 所示。

职工基本信息源数据文件类型为以英文分号为分隔符的文本文件，该数据有数据不标准，格式不符合要求等问题，要求利用 Kettle 软件对该数据进行相应的清洗并且导出为 xml 数据类型进行保存维护。

采集的唱片信息数据源数据文件类型为以 xml，该数据有数据文件类型，格式不符合要求等问题，要求利用 Kettle 软件对该数据进行相应的清洗并且导出成 json 文件类型。

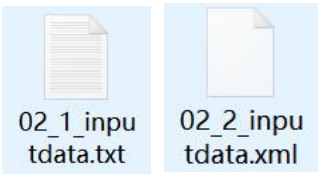
表 1 员工基本信息数据字段说明

字段名称	位置	说明	示例
Name	1	员工姓名	peter
Age	2	员工年龄	24
Birth	3	员工出生年月日	19950504
Sex	4	员工性别	M
Salary	5	员工薪资	5000

表 2 唱片信息数据字段说明

字段名称	位置	说明	示例
TITLE	1	唱片标题	Empire Burlesque
ARTIST	2	艺术家	Bob Dylan
COUNTRY	3	发行国家	usa
COMPANY	4	唱片公司	Columbia
PRICE	5	唱片价钱	10.90
YEAR	6	唱片发行年份	1985

源数据文档名称：02_1_inputdata.txt、02_2_inputdata.xml



以下任务需要按照步骤进行截图，所有截图保存到物理机上指定位置“E:\技能抽查提交资料\考生学校+考生号+考生姓名\T02 答案.docx”，具体操作见提交要求。

任务 1 职工基本信息数据的字段的操作

任务 1.1：转换文件的新建与数据的导入（5 分）

1. 在 Kettle 软件中新建名为 02_1 的转换文件（.ktr），选择输入类别中的文本文件输入步骤。

2. 配置文本文件输入步骤中的相关参数，将 02_1_inputdata.txt 中的源数据全部导入 kettle 中。

任务 1.2：数据的清洗（30 分）

1. 选择转换类别中的字段选择步骤，建立文本文件输入步骤到字段选择步骤之间的连接，进行以下任务：

- 删除“age”字段。
- 将“Birth”字段的数据格式改成“1995-05-04”的形式。
- 将“Sex”字段名改成“Gender”，并且移到“Name”字段后面。
- 将“Salary”字段的数据类型改成浮点型数据并且保留两位小数。

2. 选择转换类别中的字符串操作步骤，建立字段选择步骤到字符串操作步骤之间的连接，进行以下任务：

- 将“Name”字段中的数据类型设置为首字母大写

任务 1.3：数据的导出和维护（5 分）

1. 选择输出类别中的 XML output 步骤，建立字符串操作步骤到 XML output 步骤之间的连接。

2. 配置 XML output 步骤中相关参数，将清洗过的数据导出名为 02_1_outputdata 的 XML 数据。

任务 2 唱片信息数据字段的操作

任务 2.1：数据的导入（5 分）

1.在 Kettle 软件中新建名为 02_2 的转换文件（.ktr），选择输入类别中的 Get data from XML 步骤。

2.配置 Get data from XML 步骤中的相关参数，将 02_2_inputdata.xml 中

的源数据全部导入 kettle 中。

任务 2.2：数据的清洗（30 分）

1. 选择转换类别中的字段选择步骤，建立 Get data from XML 步骤到字段选择步骤之间的连接，将“TITLE”字段改为“唱片名称”，“ARTIST”改为“艺术家”，“COUNTRY”改为“国家”，“COMPANY”字段改为“公司”，“PRICE”改为“美元”，“YEAR”改为“年份”。

2. 选择转换类别中的字符串操作步骤，建立字段选择步骤到字符串操作步骤之间的连接，将“国家”字段数据类型设置为全大写形式。

3. 选择脚本类别中的公式步骤，建立字符串操作步骤到公式步骤之间的连接，新建“人民币”字段，该字段的值为“美元”字段数据乘以 6.97，且数据类型为整型。

4. 选择转换类别中的排序记录步骤，建立公式步骤到排序记录步骤之间的连接，设置按照“年份”字段升序排序。

任务 2.3：数据的导出和维护（5 分）

1.选择输出类别中的 JSON output 步骤，建立排序记录步骤到 JSON output 步骤之间的连接。

2.配置 JSON output 步骤中相关参数，将清洗过的数据导出为一个名为 02_2_outputdata 的 JSON 数据文件类型。

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：湖南信息职业技术学院 01 张三，并且在考生文件夹中创建“T02 答案.docx”文件。

2）“技能抽查提交资料”文件夹内保存截图 word 文档、转换源文件及引用的相关数据文件，转换源文件以“姓名_题号.ktr”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

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 8GB 以上,硬盘 100G	无

②软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7	安装 64 位版本
2	Kettle	12.0 及以上	导入 mysql 的 jar 包
3	Mysql	5.7	3. 开放进行外部连接权限 4. 时区设置为“+8:00”
4	Navicat for MySQL	11.0 及以上	无
5	Microsoft Office	2007 及以上	无

（3）考核时量

考核时间为 150 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：

表 3 数据清洗模块考核评价标准

评价内容			配分	评分标准		备注
工作任务	模块一	数据的导入	5 分	数据输入步骤的选择符合要求	2 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。
				源数据按照要求导入kettle中	3 分	
		数据的清洗	30 分	字段选择步骤的选择符合要求	2 分	
				删除“age”字段	5 分	
				按照要求修改“Birth”字段的数据格式	5 分	
				按照要求对“Sex”字段重命名	5 分	
				按照要求修改“Salary”字段的数据类型	5 分	
				字符串操作步骤的选择和连接符合要求	3 分	
				按照要求修改将“Name”字段	5 分	
		数据的导出和维护	5 分	XML output步骤的选择和连接符合要求	2 分	
				数据正确导出为xml文件	3 分	
	模块二	数据的导入	5 分	数据输入步骤的选择符合要求	2 分	

				源数据按照要求导入kettle中	3分	
		数据的清洗	30 分	字段选择步骤的选择和连接符合要求	2 分	
				按照要求修改字段名称	8分	
				字符串操作步骤的选择和连接符合要求	2分	
				按照要求设置“国家”字段数据类型	3分	
				公式步骤的选择和连接符合要求	2分	
				公式步骤条件设置符合要求	8分	
				排序记录步骤的选择和连接符合要求	2分	
				设置按照“年份”字段升序排序	3分	
		数据的导出和维护	5 分	数据输出步骤符合要求	2 分	
				数据按照要求正确导出为JSON文件	3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			

28.试题编号：3-1-3：csv 数据清洗

（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）和订单详情数据（03_2_inputdata.csv）的各字段说明如下表 1 和表 2 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
ordertype	9	订单类型	SHOPPING
payexceptiontype	11	订单异常菜单列表	COMMON
boxid	12	售货机 ID	73216297342
payedtime	13	支付时间	2018/11/21 18:46:14
ordernum	15	订单号	272074322789605000

表 2 无人售客户订单详情数据字段说明

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

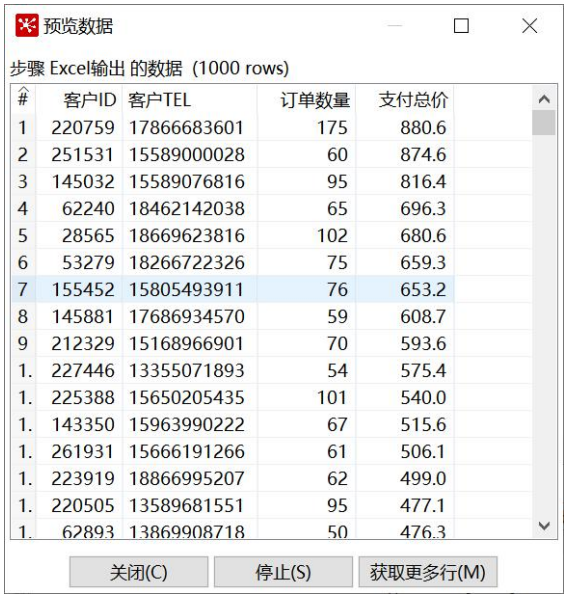


图 1 客户订单聚合结果预览

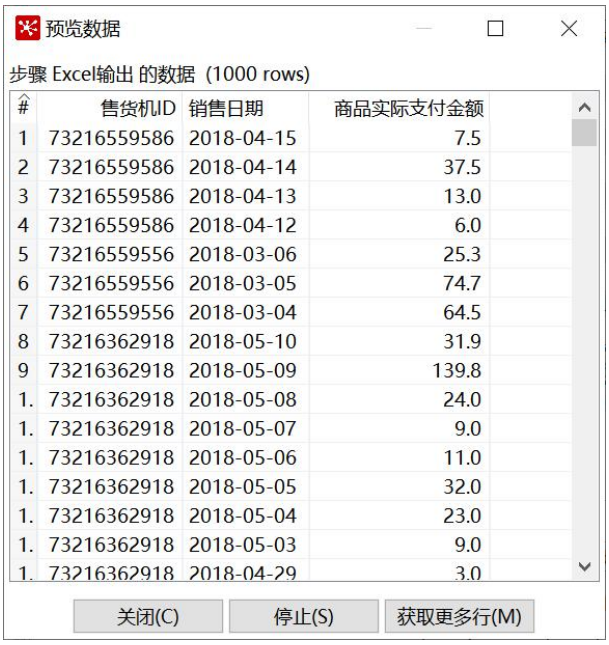
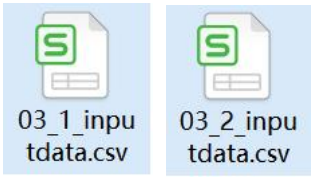


图 2 售货机日销售金额预览

源数据文档名称：03_1_inputdata.csv、03_2_inputdata.csv



以下任务需要按照步骤进行截图，所有截图保存到物理机上指定位置“E:\技能抽查提交资料\考生学校+考生号+考生姓名\T03 答案.docx”，具体操作见提交要求。

任务 1 客户订单信息表数据的分组聚合操作

任务 1.1：数据的导入（5 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。

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。

任务 1.2：数据的清洗（30 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 1.3：数据的导出和维护（5 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。

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。

任务 2 各售货机日销金额统计操作

任务 2.1：数据的导入（5 分）

1.在 Kettle 软件中新建名为 03_2 的转换文件（.ktr），选择输入类别中的 CSV 文件输入步骤。

2.配置 CSV 文件输入步骤中的相关参数，将 03_2_inputdata.csv 中的源数据全部导入 kettle 中,要求将 createdtime 字段的字段类型设置为 String。

任务 2.2：数据的清洗（30 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，过滤得到商品名称不为空和支付成功的订单数据。

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 boxid、createdtime、amount、e 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。

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。

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。

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。

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。

任务 2.3：数据的导出和维护（5 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，

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_2_outputdata.xlsx 的数据。

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：湖南信息职业技术学院 01 张三，并且在考生文件夹中创建“T03 答案.docx”文件。

2）“技能抽查提交资料”文件夹内保存截图 word 文档、转换源文件及引用的相关数据文件，转换源文件以“姓名_题号.ktr”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

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 8GB 以上,硬盘 100G	无

②软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7	安装 64 位版本
2	Kettle	12.0 及以上	导入 mysql 的 jar 包
3	Mysql	5.7	5. 开放进行外部连接权限 6. 时区设置为“+8:00”
4	Navicat for MySQL	11.0 及以上	无
5	Microsoft Office	2007 及以上	无

（3）考核时量

考核时间为 150 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：

表 3 数据清洗模块考核评价标准

评价内容			配分	评分标准		备注
工作任务	模块一	数据的导入	5 分	数据输入步骤的选择符合要求	2 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。
				源数据按照要求导入kettle中	3 分	
		数据的清洗	30 分	按照要求选择和设置过滤记录步骤的参数	6 分	
				按照要求选择和设置过字段选择步骤的参数	6 分	
				按照要求选择和设置过排序记录步骤的参数	6 分	
				按照要求选择和设置分组聚合步骤的参数	6 分	
				按照要求选择和设置过排序记录步骤的参数	6 分	
		数据的导出和维护	5 分	Microsoft Excel 输出的选择和连接符合要求	2 分	
				处理过的数据正确导出为xlsx文件数据	3 分	

	模块二	数据的导入	5 分	数据输入步骤的选择符合要求	2 分	
				源数据按照要求导入kettle中	3分	
		数据的清洗	30 分	按照要求选择和设置过滤记录步骤的参数	5 分	
				按照要求选择和设置字段选择步骤的参数	5分	
				按照要求选择和设置剪切字符串步骤的参数	5分	
				按照要求选择和设置过排序记录步骤的参数	5分	
				按照要求选择和设置分组聚合步骤的参数	5分	
				按照要求选择和设置过排序记录步骤的参数	5分	
		数据的导出和维护	5 分	数据输出步骤符合要求	2 分	
				数据按照要求正确导出为JSON文件	3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			

29. 试题编号：3-1-4，JS 数据清洗

（1）任务描述

某企业采集的相关电影信息数据各字段说明如下表 1 所示，源数据文件类型为 JSON，该数据有数据文件类型，格式不符合要求等问题，要求利用 Kettle 对该数据先进行进行清洗和校验，最后将处理过的数据导出成 CSV 文件类型，数据预览结果如下图 1 所示。

表 1 员工信息表

字段名称	位置	说明	示例
主演	1	影片主演	主演：张国荣,张丰毅,巩俐
上映时间	2	影片上映时间	上映时间：1993-07-26
评分	3	影片综合评分	9.5
片名	4	影片名字	霸王别姬
宣传图片	5	宣传图片网址	https:\/\/p0.meituan.net\/movie\/ce4da3e03e655b5b88ed31b5cd7896cf62472.jpg@160w_220h_1e_1c
编号	6	影片编号	1

#	编号	片名	评分	第一主演	上映年份
1	1	霸王别姬	9.5	张国荣	1993
2	2	肖申克的救赎	9.5	蒂姆·罗宾斯	1994
3	3	这个杀手不太冷	9.5	让·雷诺	1994
4	4	罗马假日	9.0	格利高里·派克	1953
5	5	泰坦尼克号	9.4	莱昂纳多·迪卡普里奥	1998
6	6	唐伯虎点秋香	9.1	周星驰	1993
7	7	乱世佳人	9.1	费雯·丽	1939
8	8	魂断蓝桥	9.2	费雯·丽	1940
9	9	辛德勒的名单	9.2	连姆·尼森	1993
1.	10	喜剧之王	9.1	周星驰	1999
1.	11	天空之城	9.0	寺田农	1992
1.	12	音乐之声	9.0	朱莉·安德鲁斯	1965
1.	13	大闹天宫	9.0	邱岳峰	1965
1.	14	春光乍泄	9.2	张国荣	1997
1.	15	剪刀手爱德华	8.8	约翰尼·德普	1990
1.	16	黑客帝国	9.0	基努·里维斯	2000

图 1 清洗过的电影数据预览

源数据文档名称：04inputdata.js

以下任务需要按照步骤进行截图，所有截图保存到物理机上指定位置“E:\技能抽查提交资料\考生学校+考生号+考生姓名\T04 答案.docx”，具体操作见提交要求。

任务 1：数据的导入和转换（10 分）

1.在 Kettle 软件中新建名为 04 的转换文件（.ktr），选择输入类别中的 JSON input 步骤。

2.配置 JSON input 步骤中的相关参数，将 04inputdata.js 中的源数据全部导入 kettle 中。

任务 2：数据的清洗，处理后的数据示例如表（50 分）

1．选择转换类别中的列拆分为多行步骤，建立 JSON input 步骤到列拆分为多行步骤之间的连接。以英文逗号‘,’作为分割符拆分‘主演’字段，新字段名设置为‘主演（new）’。

2．选择流程类别中的过滤记录步骤，建立列拆分为多行步骤到过滤记录步骤之间的连接。设置过滤条件为：‘主演 new’字段 STARTS WITH 主演(String)，过滤得到‘主演 new’字段中的主演名字排名在第一位的主演。

3．选择转换类别中的剪切字符串步骤，建立列过滤记录步骤到剪切字符串步骤之间的连接。设置过滤记录步骤，将发送 true 数据给步骤：剪切字符串。

4．设置剪切字符串步骤，剪切时间，输入流字段为：上映时间，输出流字段为：上映年份，起始位置：5，结束位置：9；剪切主演，输入流字段：主演 new，输出流字段：第一主演，起始位置：3，结束位置：20。

5．对清洗处理的数据进行抽取。选择转换类别中的字段选择步骤，建立过剪切字符串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按照一定的顺序，要求字段为编号，片名，评分，第一主演，上映年份的顺序。

任务 3：数据的导出和维护（20 分）

1.选择输出类别中的 Microsoft Excel 输出步骤，建立字段选择步骤到 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，04_1_outputdata.xlsx 数据文件类型。

3.选择输出类别中的 XML output 步骤，建立字段选择步骤到 XML output 步骤之间的连接。

4.配置 XML output 步骤中相关参数，将清洗过的数据导出名为 04_2_outputdata 的 XML 数据。

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：湖南信息职业技术学院 01 张三，并且在考生文件夹中创建“T04 答案.docx”文件。

2）“技能抽查提交资料”文件夹内保存截图 word 文档、转换源文件及引用的相关数据文件，转换源文件以“姓名_题号.ktr”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

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 8GB 以上,硬盘 100G	无

②软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7	安装 64 位版本
2	Kettle	12.0 及以上	导入 mysql 的 jar 包
3	Mysql	5.7	7. 开放进行外部连接权限 8. 时区设置为“+8:00”
4	Navicat for MySQL	11.0 及以上	无
5	Microsoft Office	2007 及以上	无

（3）考核时量

考核时间为 150 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：

表 2 数据清洗模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	数据的导入	10 分	数据输入步骤的选择是否符合要求	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。
			参数的配置符合要求,并且源数据按照要求导入 kettle 中	5 分	
	数据的清洗	56 分	列拆分为多行步骤选择和连接符合要求	2 分	
			“主演”字段按照要求进行拆分	10 分	
			过滤记录步骤选择正确	2 分	
			过滤记录步骤的过滤条件设置符合要求	10 分	
			剪切字符串步骤选择和连接正确	5 分	
			过滤记录步骤与剪切字符串步骤连接正确	5 分	
			按照要求正确剪切‘上映时间’和‘主演 new’字段	10 分	
			字段选择步骤的选择和连接正确	2 分	
			按照要求进行数据进行抽取	10 分	
	数据的导出和维护	14 分	Microsoft Excel 输出步骤的选择和连接符合要求	2 分	
			处理过的数据正确导出为 xlsx文件数据	5 分	
			XML output步骤的选择和连接符合要求	2 分	
			处理过的数据正确导出为XML文件数据	5 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			

30.试题编号：3-1-5：CSV 数据综合清洗

（1）任务描述

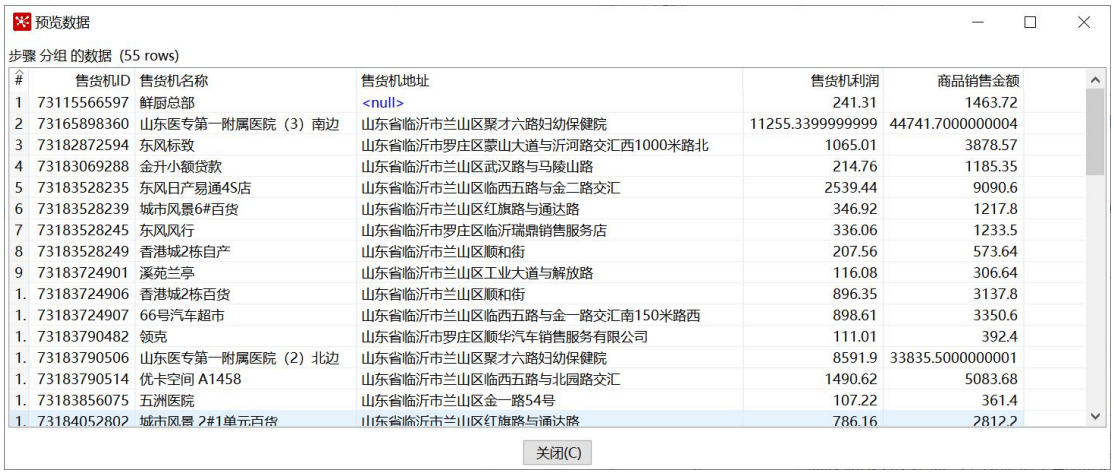
某无人售货机公司采集无人售货机客户订单详情信息数据（05_1_inputdata.csv）和无人售货机信息数据（05_2_inputdata.csv）各字段说明如下表 1 和表 2 所示，客户订单详情数据记录着不同客户每天的每一笔订单详细数据，无人售货机信息数据记录着售货机的基本信息，从盈利的角度出发，为了了解售货机每天销售的情况，统计每台机器的利润，要求利用 Kettle 软件对该数据进行过滤、公式计算、排序等清洗操作得到每台售货机每天的利润，清洗过后的数据如下图 1 所示。

表 1 无人售客户订单详情数据字段说明

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

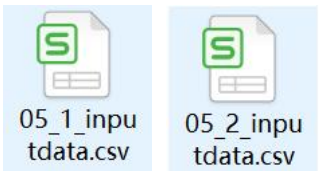
字段名称	位置	说明	示例
boxid	1	售货机 ID	73216297360
address	2	售货机投放地址	山东省临沂市兰山区红旗路与通达路
name	3	售货机名称	城市风景 2#2 单元自产
qrcode	4	售货机二维码	http://assets.mayihezi.com/prod/box/Qrcode/20189/20193050656de5cfc48-6ba2-4bac-8997-42455437bde2.jpg
serialnumber	5	售货机编码	A2448
status	6	售货机状态	ONLINE
modelnumber	7	售货机型号	MA650-A



#	售货机ID	售货机名称	售货机地址	售货机利润	商品销售金额
1	73115566597	鲜厨总部	<null>	241.31	1463.72
2	73165898360	山东医专第一附属医院（3）南边	山东省临沂市兰山区聚才六路妇幼保健院	11255.3399999999	44741.7000000004
3	73182872594	东风标致	山东省临沂市罗庄区蒙山大道与沂河路交汇西1000米路北	1065.01	3878.57
4	73183069288	金升小额贷款	山东省临沂市兰山区武汉路与马陵山路	214.76	1185.35
5	73183528235	东风日产易通4S店	山东省临沂市兰山区临西五路与金二路交汇	2539.44	9090.6
6	73183528239	城市风景6#百货	山东省临沂市兰山区红旗路与通达路	346.92	1217.8
7	73183528245	东风风行	山东省临沂市罗庄区临沂瑞鼎销售服务店	336.06	1233.5
8	73183528249	香港城2栋自产	山东省临沂市兰山区顺和街	207.56	573.64
9	73183724901	溪苑兰亭	山东省临沂市兰山区工业大道与解放路	116.08	306.64
1.	73183724906	香港城2栋百货	山东省临沂市兰山区顺和街	896.35	3137.8
1.	73183724907	66号汽车超市	山东省临沂市兰山区临西五路与金一路交汇南150米路西	898.61	3350.6
1.	73183790482	领克	山东省临沂市罗庄区顺华汽车销售服务有限公司	111.01	392.4
1.	73183790506	山东医专第一附属医院（2）北边	山东省临沂市兰山区聚才六路妇幼保健院	8591.9	33835.5000000001
1.	73183790514	优卡空间 A1458	山东省临沂市兰山区临西五路与北园路交汇	1490.62	5083.68
1.	73183856075	五洲医院	山东省临沂市兰山区金一路54号	107.22	361.4
1.	73184052802	城市风景 2#1单元百货	山东省临沂市兰山区红旗路与通达路	786.16	2812.2

图 1 售货机利润预览图

源数据文档名称：05_1_inputdata.csv、05_2_inputdata.csv



以下任务需要按照步骤进行截图，所有截图保存到物理机上指定位置“E:\

技能抽查提交资料\考生学校+考生号+考生姓名\T05 答案.docx”，具体操作见提交要求。

任务 1：数据的导入和转换（5 分）

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

2.配置 CSV 文件输入步骤中的相关参数，将 05_1_inputdata.csv 中的源数据全部导入 kettle 平台中。

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 05_2_inputdata.csv 中的源数据全部导入 kettle 平台中。

任务 2：数据的清洗，处理后的数据示例如表（65 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。

6．对订单数据按照售货机 ID 进行排序，选择转换类别中的排序记录步骤，建立公式步骤到排序记录步骤之间的连接，设置按照“售货机 ID”字段升序排序。

7．对售货机的商品利润等字段进行分组聚合，统计各个售货机的利润。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、售货机名称和售货机地址，聚合名称一为“售货机利润”，是商品利润求和得到，名称二为“商品销售金额”，是商品实际支付总金额求和得到。

任务 3：数据的导出和维护（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 05_1_outputdata 的 xlsx 的数据。

3.选择输出类别中的 JSON output 步骤，建立分组聚合步骤到 JSON output 步骤之间的连接。

4.配置 JSON output 步骤中相关参数，将清洗过的数据导出为一个名为 05_2_outputdata 的 JSON 数据文件类型。

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：湖南信息职业技术学院 01 张三，并且在考生文件夹中创建“T05 答案.docx”文件。

2）“技能抽查提交资料”文件夹内保存截图 word 文档、转换源文件及引用的相关数据文件，转换源文件以“姓名_题号.ktr”命名，最终将考生文件夹进行压缩后提交。

（2）实施条件

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 8GB 以上,硬盘 100G	无

②软件环境

序号	软件	版本	备注
1	桌面版操作系统	Windows 7	安装 64 位版本
2	Kettle	12.0 及以上	导入 mysql 的 jar 包
3	Mysql	5.7	9. 开放进行外部连接权限 10. 时区设置为“+8:00”
4	Navicat for MySQL	11.0 及以上	无
5	Microsoft Office	2007 及以上	无

（3）考核时量

考核时间为 150 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：

表 3 数据清洗模块考核评价标准

评价内容			配分	评分标准		备注
工作任务	模块一	数据的导入	5 分	数据输入步骤的选择符合要求	2 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。
				两份源数据按照要求导入 kettle中	3 分	
		数据的清洗	65 分	按照要求选择和设置过滤记录步骤的参数	10 分	

				按照要求选择过字段选择步骤并且设置正确的参数	8 分	2、严重违反考场纪律、造成恶劣影响的本项目记0分。
				按照要求选择记录关联（笛卡尔输出)步骤并且设置正确的参数	8 分	
				按照要求选择过字段选择步骤并且设置正确的参数	10 分	
				按照要求选择公式组件步骤并且设置正确的参数	11 分	
				按照要求选择排序步骤并且设置正确的参数	8 分	
				按照要求选择分组聚合步骤并且设置正确的参数	10 分	
		数据的导出和维护	10 分	Microsoft Excel 输出步骤的选择和连接符合要求	2 分	
				处理过的数据正确导出为xlsx文件数据	3 分	
				JSON output步骤的选择和连接符合要求	2 分	
				处理过的数据正确导出为JSON文件数据	3 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			

项目 2　Spark 大数据处理与分析

31 试题编号：3-2-1：Spark 大数据分析平台搭建

（1）任务描述

某大型互联网公司的业务生产系统数据规模不断增加，每天产生海量的生产数据，这些数据既包括文本、文档、图片、视频等非结构化的数据，同时又包括生产系统和业务系统的结构化数据。为了公司生产系统安全高可用，同时能够统

一存储、收集、管理、分析和挖掘这些海量数据，为了构建实时计算，快速计算海量 Spark 大数据平台，为实现实时数据可视化系统，推荐系统，提供中台支持、促进信息技术和数据资源充分利用。该公司拟搭建 Spark 大数据分析平台，提供快速数据分析、数据挖掘和快速检索等功能。

经公司 CIO 反复调研，决定选用开源 Spark 构建大数据平台和大数据系统应用研发。搭建 Spark 分布式大数据平台，以实现对大数据分析与管理、为智能推荐，智能检索等，提供统一数据中台支持。

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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一 基本环境配置（40 分）

本次部署三个服务器节点，节点主机名和 IP 如下表所示：

表 1-主机名和 IP 地址表

主机名	IP 地址	备注
master	192.168.15.X	服务器节点、主、从节点
slave01	192.168.15.X	服务器节点、从节点
slave02	192.168.15.X	服务器节点、从节点

1、根据表 1，设置三个服务节点的主机名为 master、slave01、slave02，将命令和结果截图后存放到文档中（图片标题为“任务一：基本环境配置-1”）。（10）

2、将三个节点的网卡设置为 NAT,手动设置 IP 地址，使三个节点能够 ping 通 www.baidu.com，将命令和结果截图后存放到文档中（图片标题为“任务一：基本环境配置-2”）。（10）

3、在 hosts 文件设置三个服务节点的主机名与 IP 映射，将命令和结果截图后存放到文档中（图片标题为“任务一：基本环境配置-3”）。（5）

4、关闭防火墙，设置开机不启动，再查看防火墙状态，将命令和结果截图后存放到文档中（图片标题为“任务一：基本环境配置-4”）。（5）

5、修改/etc/selinux/config 文件，将原来的 SELINUX=enforcing 修改为 SELINUX=permissive，再用 setenforce 0 命令将当前的 SELinux 模式设置为 permissive，用 getenforce 命令查看结果，将命令和结果截图后存放到文档中（图片标题为“任务一：基本环境配置-4”）。（10）

任务二 安装 JDK（40 分）

1、查询是否安装 JDK 软件。将命令和结果截图后存放到文档中（图片标题

为“任务二：安装 JDK-1”）（3 分）

2、如果安装的版本低于 1.7 或者是 centos7 自带的 OpenJDK，则卸载该 JDK。将命令和结果截图后存放到文档中（图片标题为“任务二：安装 JDK-2”）（2 分）

3、查看 JDK 安装路径。将命令和结果截图后存放到文档中（图片标题为“任务二：安装 JDK-3”）（5 分）

4、用 SecureCRT 或者 Xshell 等工具将 JDK 导入到 opt 目录下面的 software 文件夹下面。将命令和结果截图后存放到文档中（图片标题为“任务二：安装 JDK-4”）（5 分）

5、解压 JDK 到/opt/module 目录下。将命令和结果截图后存放到文档中（图片标题为“任务二：安装 JDK-5”）（5 分）

6、配置 JDK 环境变量。将命令和结果截图后存放到文档中（图片标题为“任务二：安装 JDK-6”）（5 分）

7、测试 JDK 是否安装成功。将命令和结果截图后存放到文档中（图片标题为“任务二：安装 JDK-7”）（5 分）

8、将 JDK 和相关配置文件分发到 slave01、slave02。（10 分）

（2）实施条件

表 3-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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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%。考试舞弊、抄袭、没有按要求填写相关信息，本项目记 0 分，严重违反考场纪律、造成恶劣影响的本项目记 0 分。具体评价标准见下面描述：

评分项一：基本环境配置

序号	评分内容	评分点	分值（分）
1	设置三个节点的主机名	正确设置主机名	10 分
2	手动设置三个节点IP地址，并能ping通	正确设置三台机器静态 IP	10 分
3	设置三个节点主机名与IP映射	正确设置主机名与 IP 映射	5 分
4	关闭防火墙，设置开机不启动	正确关闭防火墙	5 分
5	修改/etc/selinux/config 文件，为 SELINUX=permissive	SELINUX 设置正确	10分

评分项二：安装 JDK

序号	评分内容	评分点	分值（分）
1	查询是否安装JDK软件	正确安装 JDK	3 分
2	卸载OpenJDK	OpenJDK 卸载	2 分
3	检查JDK卸载是否成功	验证是的卸载成功	5 分
4	将JDK导入到opt目录下	上传 JDK	5 分
5	解压JDK到/opt/module下	解压是否成功	5 分
6	配置JDK环境变量	配置 JDK 环境变量	5 分
7	测试JDK是否安装成功	测试 JDK	5 分
8	将JDK和相关配置文件分发到slave01、slave02	分发配置文件	10 分

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	基本环境配置	40 分	设置三个节点的主机名	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
			手动设置三个节点IP地址，并能ping通	10 分	
			设置三个节点主机名与IP映射	5 分	
			关闭防火墙，设置开机不启动	5 分	
			修改/etc/selinux/config 文件，为 SELINUX=permissive	10分	
	安装JDK	40 分	查询是否安装JDK软件	3 分	
			卸载OpenJDK	2 分	

			检查JDK卸载是否成功	5 分	记0分。
			将JDK导入到opt目录下	5 分	
			解压JDK到/opt/module下	5 分	
			配置JDK环境变量	5 分	
			测试JDK是否安装成功	5 分	
			将JDK和相关配置文件分发到slave01、slave02	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			

32 试题编号：3-2-2：Spark 大数据分析平台搭建

（1）任务描述

某大型互联网公司的业务生产系统数据规模不断增加，每天产生海量的生产数据，这些数据既包括文本、文档、图片、视频等非结构化的数据，同时又包括生产系统和业务系统的结构化数据。为了公司生产系统安全高可用，同时能够统一存储、收集、管理、分析和挖掘这些海量数据，为了构建实时计算，快速计算海量 Spark 大数据平台，为实现实时数据可视化系统，推荐系统，提供中台支持、促进信息技术和数据资源充分利用。该公司拟搭建 Spark 大数据分析平台，提供快速数据分析、数据挖掘和快速检索等功能。

经公司 CIO 反复调研，决定选用开源 Spark 构建大数据平台和大数据系统应用研发。搭建 Spark 分布式大数据平台，以实现对大数据分析与管理、为智能推荐，智能检索等，提供统一数据中台支持。

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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一 ZooKeeper 组件安装（40 分）

本次部署三个服务器节点，节点主机名和 IP 如下表所示：

表 1 主机名和 IP 地址表

主机名	IP 地址	备注
master	192.168.15.X	服务器节点、主、从节点
slave01	192.168.15.X	服务器节点、从节点
slave02	192.168.15.X	服务器节点、从节点

1、根据表集群规划，在 master、slave01 和 slave02 三个节点上部署 Zookeep 将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 ZooKeeper-1”）(5 分)

2、解压 Zookeeper 安装包到/opt/module/目录下，scp 命令同步分发 /opt/module/zookeeper-3.4.10 目录内容到 slave01、slave02。 将命令和结果

截图后存放到文档中（图片标题为“任务 1：安装 ZooKeeper-2”）（5 分）

3、配置服务器编号在/opt/module/zookeeper-3.4.10/这个目录下创建 zkData，

并创建一个 myid 的文件，编辑 myid 文件，在文件中添加与 server 对应的编号。将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 ZooKeeper-3”）（10 分）

4、最后拷贝配置好的 zookeeper 到其他机器上，并分别在 master、slave01、slave02 上修改 myid 文件中内容为 3、4。将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 ZooKeeper-4”）（10 分）

5、配置 zoo.cfg 文件，重命名/opt/module/zookeeper-3.4.10/conf 这个目录下的

zoo_sample.cfg 为 zoo.cfg，修改 zoo.cfg，同步 zoo.cfg 配置文件。将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 ZooKeeper-4”）（10 分）

任务二　安装 Spark（40 分）

1、解压 Spark 安装包，进入 Spark 安装目录下的 conf 文件夹，修改配置文件名称。将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 spark-1”）（10 分）

2、修改 slave 文件，添加 work 节点。将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 spark-2”）（10 分）

3、修改 spark-env.sh 文件，添加配置。将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 spark-3”）（10 分）

4、分发 spark 安装包，并且启动 spark。将命令和结果截图后存放到文档中（图片标题为“任务 1：安装 spark-4”）（10 分）

（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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版

4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：ZooKeeper 组件安装

序号	评分内容	评分点	分值（分）
1	部署Zookeeper	三个节点上部署Zookeeper	5 分
2	解压Zookeeper安装包到/opt/module/目录下	正确解压Zookeeper安装包	5 分
3	配置服务器编号	正确配置服务器编号	10 分
4	配置好zookeeper	拷贝配置好的zookeeper到其他机器，并修改编号	10 分
5	配置zoo.cfg	正确配置zoo.cfg	10分

评分项二：安装 Spark

序号	评分内容	评分点	分值（分）
1	解压Spark安装包，修改配置文件名称	正确解压安装包	10 分
2	修改slave文件，添加work节点	正确配置 slave 文件	10 分
3	修改spark-env.sh文件，添加配置。	正确配置 spark-env.sh 文件	10 分
4	分发spark安装包，并且启动spark	正确分发 spark 安装包，并成功启动 spark	10 分

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	ZooKeeper组件安装	40 分	三个节点上部署Zookeep	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
			解压Zookeeper安装包到/opt/module/目录下	5 分	
			配置服务器编号	10 分	
			拷贝配置好的zookeeper到其他机器，并修改编号	10 分	
			正确配置zoo.cfg	10分	

	安装Spark	40 分	解压Spark安装包，修改配置文件名称	10 分	考场纪律、造成恶劣影响的本项目记0分。
			修改slave文件，添加work节点	10 分	
			修改spark-env.sh文件，添加配置。	10 分	
			分发spark安装包，并且启动spark	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：3-2-3：Spark 开发-网站用户访问日志数据分析

（1）任务描述

某竞赛网站每年开展数据挖掘的竞赛，在竞赛期间网站会有大量人群访问，生成了大量用户访问记录。现提供部分脱敏访问日志数据。日志数据的基本内容如下表所示。

要求：根据提供的用户访问日志数据，利用 Spark 技术对日志数据探索与分析

表 1 字段信息表

属性名称	属性解释
Id	序号
Content_id	网页 ID
page_path	网址
Userid	用户 ID
Sessionid	缓存生成 ID
Data_time	访问时间

```
1   478896,1043,/jszx/1043.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:23:07
2   478897,983,/news/983.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:23:12
3   478900,983,/news/983.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:24:35
4   478901,1043,/jszx/1043.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:24:36
5   478903,654,/tj/654.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:24:51
6   478913,654,/tj/654.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:26:42
7   478923,747,/tj/747.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:27:48
8   478926,661,/tj/661.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:28:30
9   478927,654,/tj/654.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:28:42
10  478932,654,/tj/654.jhtml,14884,0A159405E855CB353D28FE77242A6629,2017-03-01 00:29:34
11  478946,654,/tj/654.jhtml,14884,850120FE530FBAE408177B45507AD647,2017-03-01 00:32:46
12  479020,661,/tj/661.jhtml,14886,991E5AE58D384A92E8AB8A5933FF819B,2017-03-01 01:04:07
13  479319,661,/tj/661.jhtml,14887,634B05DF4A8BA61886183249CF79699A,2017-03-01 03:52:28
14  479743,661,/ts/661.jhtml,14888,A07CA0F6BB1EC3AE1169410195D7813F,2017-03-01 08:23:20
15  479761,661,/tj/661.jhtml,14888,A07CA0F6BB1EC3AE1169410195D7813F,2017-03-01 08:27:12
16  479774,578,/tj/578.jhtml,14888,A07CA0F6BB1EC3AE1169410195D7813F,2017-03-01 08:28:39
17  479777,654,/tj/654.jhtml,14889,92ACA815DC1C1786F2D5FFB47D2DC97D,2017-03-01 08:28:53
18  479782,654,/tj/654.jhtml,14889,92ACA815DC1C1786F2D5FFB47D2DC97D,2017-03-01 08:29:47
19  479850,1030,/jingsa/1030.jhtml,14890,50A80AE039DDC4FFEEA03678D65D0779,2017-03-01 08:52:04
20  479856,1030,/jingsa/1030.jhtml,14890,50A80AE039DDC4FFEEA03678D65D0779,2017-03-01 08:52:24
```

图 1 日志数据部分展示

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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：配置 Spark 的 IDEA 开发环境。（20 分）

1、IDEA 中 Scala 插件安装。

2、Maven 软件安装及配置。

3、IDEA 中正确配置 Maven 工程

任务二：对访问记录中的网页去重，统计本周期内被访问的网页的个数。（20 分）

1、完成统计本周期内被访问的网页的个数程序编写。

2、打包程序，编译为“count.jar”的 JAR 包。

3、上传 count.jar 到 Linux 下的/opt 目录中，进入 Linux 的 Spark 安装包下的 bin 目录，执行以下命令。./spark-submit --master yarn-cluster --class demo.UserCount /opt/count.jar /user/root/jc_content_viewlog.txt web_countFile user_countFile date_countFile。

任务三：对 userid 去重，统计登录用户的数量。（20 分）

1、完成对 userid 去重，统计登录用户的数量程序编写。

2、打包程序，编译为“count.jar”的 JAR 包。

3、上传 count.jar 到 Linux 下的/opt 目录中，进入 Linux 的 Spark 安装包下的 bin 目录，执行以下命令。./spark-submit --master yarn-cluster --class demo.UserCount /opt/count.jar /user/root/jc_content_viewlog.txt web_countFile user_countFile date_countFile。

任务四：按日统计访问记录数。（20 分）

1、完成按日统计访问记录程序编写。

2、打包程序，编译为“count.jar”的 JAR 包。

3、上传 count.jar 到 Linux 下的/opt 目录中，进入 Linux 的 Spark 安装包下的 bin 目录，执行以下命令。./spark-submit --master yarn-cluster --class demo.UserCount /opt/count.jar /user/root/jc_content_viewlog.txt web_countFile user_countFile date_countFile。

（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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：配置 Spark 的 IDEA 开发环境

序号	评分内容	评分点	分值（分）
1	IDEA中Scala插件安装	正确安装插件	5 分
2	Maven软件安装及配置	正确安装和配置maven	5 分
3	IDEA中正确配置Maven工程	正确配置maven工程	10 分

评分项二：统计本周期内被访问的网页的个数

序号	评分内容	评分点	分值（分）
1	完成统计本周期内被访问的网页的个数程序编写。	正确编写代码	10 分
2	打包程序，编译为JAR包。	打包成功	5 分
3	上传 count.jar 到 Linux 下的/opt目录中，进入Linux的Spark安装包下的bin目录，提交jar运行。	上传 jar 包，并执行成功	5 分

评分项三：统计登录用户的数量

序号	评分内容	评分点	分值（分）
1	完成对userid去重，统计登录用户的数量程序编写。	正确编写代码	10 分
2	打包程序，编译为JAR包。	打包成功	5 分

3	上传count.jar到Linux下的/opt目录中，进入Linux的Spark安装包下的bin目录，提交jar运行。	上传jar包，并执行成功	5分

评分项四：按日统计访问记录数

序号	评分内容	评分点	分值（分）
1	完成按日统计访问记录程序编写。	正确编写代码	10分
2	打包程序，编译为JAR包。	打包成功	5分
3	上传count.jar到Linux下的/opt目录中，进入Linux的Spark安装包下的bin目录，提交jar运行。	上传jar包，并执行成功	5分

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	配置Spark的IDEA开发环境	20分	1、IDEA中Scala插件安装。 2、Maven软件安装及配置。 3、IDEA中正确配置Maven工程	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。
	统计本周期内被访问的网页的个数	20分	1、完成统计本周期内被访问的网页的个数程序编写。 2、打包程序，编译为JAR包。 3、上传count.jar到Linux下的/opt目录中，进入Linux的Spark安装包下的bin目录，提交jar运行。	20分	
	统计登录用户的数量	20分	1、完成对userid去重，统计登录用户的数量程序编写。 2、打包程序，编译为JAR包。 3、上传count.jar到Linux下的/opt目录中，进入Linux的Spark安装包下的bin目录，提交jar运行。	20分	
	按日统计访问记录数	20分	1、完成按月统计访问记录程序编写。 2、打包程序，编译为JAR包。 3、上传count.jar到Linux下的/opt目录中，进入Linux的Spark安装包下的bin目录，提交jar运行。	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	0-10	

			范，能做到见名知意，需要一定的注释进行解释说明。	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			

34 试题编号：3-2-4：Spark 开发-网站用户访问日志数据分析

（1）任务描述

某竞赛网站每年开展数据挖掘的竞赛，在竞赛期间网站会有大量人群访问，生成了大量用户访问记录。现提供部分脱敏访问日志数据。日志数据的基本内容如下表所示。

要求：根据提供的用户访问日志数据，利用 Spark 技术对日志数据探索与分析

表 1 字段信息表

属性名称	属性解释
Id	序号
Content_id	网页 ID
page_path	网址
Userid	用户 ID
Sessionid	缓存生成 ID
Data_time	访问时间

```
478896,1043,/jszx/1043.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:23:07
478897,983,/news/983.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:23:12
478900,983,/news/983.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:24:35
478901,1043,/jszx/1043.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:24:36
478903,654,/tj/654.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:24:51
478913,654,/tj/654.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:26:42
478923,747,/tj/747.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:27:48
478926,661,/tj/661.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:28:30
478927,654,/tj/654.jhtml,14884,F6D362B9AFAC436D153B7084EF3BA332,2017-03-01 00:28:42
478932,654,/tj/654.jhtml,14884,0A159405E855CB353D28FE77242A6629,2017-03-01 00:29:34
478946,654,/tj/654.jhtml,14884,850120FE530FBAE408177B45507AD647,2017-03-01 00:32:46
479020,661,/tj/661.jhtml,14886,991E5AE58D384A92E8AB8A5933FF819B,2017-03-01 01:04:07
479319,661,/tj/661.jhtml,14887,634B05DF4A8BA61886183249CF79699A,2017-03-01 03:52:28
479743,661,/ts/661.jhtml,14888,A07CA0F6BB1EC3AE1169410195D7813F,2017-03-01 08:23:20
479761,661,/tj/661.jhtml,14888,A07CA0F6BB1EC3AE1169410195D7813F,2017-03-01 08:27:12
479774,578,/tj/578.jhtml,14888,A07CA0F6BB1EC3AE1169410195D7813F,2017-03-01 08:28:39
479777,654,/tj/654.jhtml,14889,92ACA815DC1C1786F2D5FFB47D2DC97D,2017-03-01 08:28:53
479782,654,/tj/654.jhtml,14889,92ACA815DC1C1786F2D5FFB47D2DC97D,2017-03-01 08:29:47
479850,1030,/jingsa/1030.jhtml,14890,50A80AE039DDC4FFEEA03678D65D0779,2017-03-01 08:52:04
479856,1030,/jingsa/1030.jhtml,14890,50A80AE039DDC4FFEEA03678D65D0779,2017-03-01 08:52:24
```

图 1 日志数据部分展示

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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：过滤出实训中访问次数在 50 次以上的用户记录并持久化到内存。（20

分）

1、上传数据到 HDFS。

2、统计并筛选数据，统计用户访问次数并筛选出访问次数在 50 次以上的用户 ID。

任务二：统计访问 50 次以上的用户主要访问的前 5 类网页。（10 分）

1、统计网页访问情况。(5 分)

任务三：合并部分网页 URL 后面带有_1、_2 字样的翻页网址，统一为一个网址。（10 分）

1、处理翻页数据，合并网页。

任务四：根据访问时间加入对应时段，6:30~11:30 为上午，11:30~14:00 为中午，14:00~17:30 为下午，17:30~19:00 为傍晚，19:00~23:00 为晚上，23:00~6:30 为深夜，统计所有用户各时段访问情况。（40 分）

1、定义一个时间段处理函数函数，用于匹配时间段并返回相应的字段值。

2、通过 map 方法对每一条记录的时间进行匹配，增加一个时间段的值到记录中。

3、将时段值作为键，值为 1，利用 reduceByKey 的方法统计各时段访问情况，最后输出统计结果。

（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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：网站用户访问日志数据分析

序号	评分内容	评分点	分值（分）
1	过滤实训中访问次数在 50 次以上的用户	程序正确过滤出实训中访问次数在 50 次以上的用户	20 分
2	统计网页访问情况	程序正确统计网页访问情况	10 分
3	处理翻页数据，合并网页	程序正确处理翻页数据，合并网页。	10 分
4	统计所有用户各时段访问情况	程序正确统计所有用户各时段访问情况，并输出统计结果。	40 分

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	过滤实训中访问次数在 50 次以上的用户	20 分	正确过滤出实训中访问次数在 50 次以上的用户	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。
	统计网页访问情况	10 分	正确统计网页访问情况	10 分	
	处理翻页数据，合并网页	10 分	正确处理翻页数据，合并网页。	10 分	
	统计所有用户各时段访问情况	40 分	正确统计所有用户各时段访问情况，并输出统计结果。	40 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：3-2-5：Spark 开发-股票数据分析

（1）任务描述

随着社会的不断发展，无论是生产者自身的资本积累，还是有限的借贷基本都难以满足企业发展的巨额资金需求。于是出现了通过发行股票来筹措资金，建立股份有限公司的办法，这就是股份制。一个人购买某个企业的股票，就成为这个公司的股东。中国大部分人对股票、股市并不陌生，投资股市称为很多人理财重要渠道，每天有数以亿人关注都市的涨幅，每天有数以百亿元的资金流入股市。

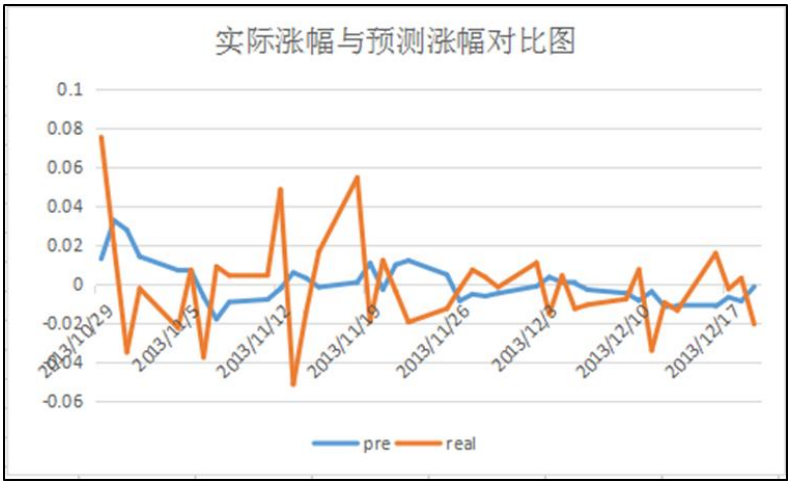
然而股价起伏难测，有些人通过炒股获得财富，也有很多人身价、财产被股市套牢。因此各界人士都开始对股价动态展开研究，分析合适股价波动的算法和模型，找出股价涨跌的规律，为很多无法根据自身能力判断股价涨跌的普通人提供指导。

本题以股价数据为基础，利用 Spark 分布式技术，完成处理股票数据处理任务，并采用移动平均实现对股价涨跌的简单预测。股票数据集包含 300 只股票的日常交易记录，记录时间长 2013 年-2016 年，采集字段总共 14 个，字段描述如下表所示

股票数据属性

属性名称	属性说明	属性名称	属性说明
Date	日期	open	开盘价
close	收盘价	low	最低价
HIigh	最高价	volume	成交量
price_change	涨跌幅	p_change	涨跌幅差值
ma5	5 日均价	v_ma5	5 日均量
ma10	10 日均价	ma20	20 日均价
v_ma10	10 日均量	v_ma20	20 日均量

股票股价预测走势和实际股价走势图



.

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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：搭建开发环境（30 分）

1、IDEA 中 Scala 插件安装。

2、Maven 软件安装及配置。

3、IDEA 中正确配置 Maven 工程

任务二：计算股价波动幅度（50 分）

1、计算股价涨跌幅公式为：

（当日收盘价-昨日收盘价）/昨日收盘价=涨跌幅 。

结果为正值，说明今日股价比昨日股价涨了，结果为负值，说明股价比昨日跌了结果为 0，则表示两日股价相同。

2、程序中对象、方法关键字大小写敏感。

3、编制程序时注意代码缩进凸显结构清晰。

4、定义类 CalculateChange 来计 算股价的涨幅。

5、运行参数配置时运行结果输出 HDFS 的路径正确。

（2）实施条件

表 3-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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：配置 Spark 的 IDEA 开发环境

序号	评分内容	评分点	分值（分）
1	IDEA中Scala插件安装	正确安装插件	10 分
2	Maven软件安装及配置	正确安装和配置maven	10 分
3	IDEA中正确配置Maven工程	正确配置maven工程	10 分

评分项二：编写计算股价波动幅度程序

序号	评分内容	评分点	分值（分）
1	计算股价涨跌幅公式为：（当日收盘价-昨日收盘价）/昨日收盘价=涨跌幅 。结果为正值，说明今日股价比昨日股价涨了，结果为负值，说明股价比昨日跌了结果为 0，则表示两日股价相同。	程序编写正确	20 分
2	程序中对象、方法关键字大小写敏感	格式正确	3 分
3	编制程序时注意代码缩进凸显结构清晰	格式正确	2 分
4	定义类 CalculateChange 来计算股价的涨幅	程序编写正确	20 分
5	运行参数配置时运行结果输出HDFS的路径正确	hdfs参数设置正确	5 分

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	搭建开发环境	30 分	1、IDEA中Scala插件安装。 2、Maven软件安装及配置。 3、IDEA中正确配置Maven工程	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
	编写计算股价波动幅度程序	50 分	1、计算股价涨跌幅公式为：（当日收盘价-昨日收盘价）/昨日收盘价=涨跌幅 。结果为正值，说明今日股价比昨日股价涨了，结果为负值，说明股价比昨日跌了结果为 0，则表示两日股价相同。 2、程序中对象、方法关键字大小	50 分	

			写敏感 3、编制程序时注意代码缩进凸显结构清晰 4、定义类 CalculateChange 来计 算股价的涨幅 5、运行参数配置时运行结果输出HDFS的路径正确		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			

36 试题编号：3-2-6：Spark 开发-股票数据分析与预测

（1）任务描述

随着社会的不断发展，无论是生产者自身的资本积累，还是有限的借贷基本都难以满足企业发展的巨额资金需求。于是出现了通过发行股票来筹措资金，建立股份有限公司的办法，这就是股份制。一个人购买某个企业的股票，就成为这个公司的股东。中国大部分人对股票、股市并不陌生，投资股市称为很多人理财重要渠道，每天有数以亿人关注都市的涨幅，每天有数以百亿元的资金流入股市。

然而股价起伏难测，有些人通过炒股获得财富，也有很多人身价、财产被股市套牢。因此各界人士都开始对股价动态展开研究，分析合适股价波动的算法和模型，找出股价涨跌的规律，为很多无法根据自身能力判断股价涨跌的普通人提供指导。

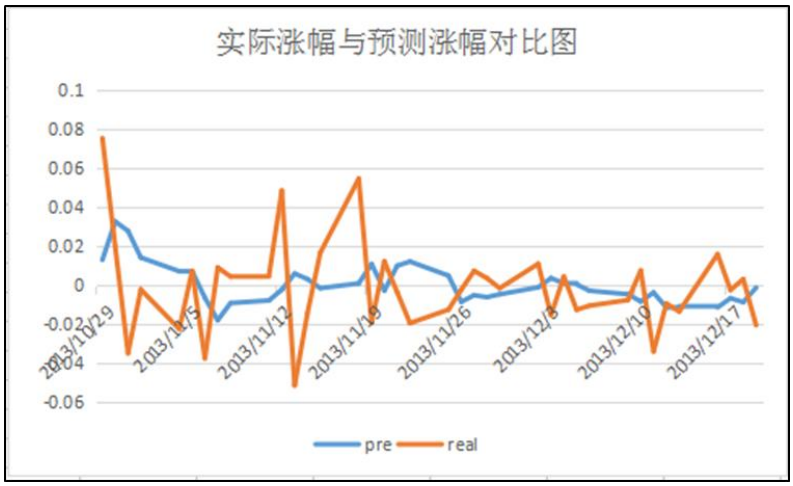
本题以股价数据为基础，利用 Spark 分布式技术，完成处理股票数据处理任务，并采用移动平均实现对股价涨跌的简单预测。股票数据集包含 300 只股票的日常交易记录，记录时间长 2013 年-2016 年，采集字段总共 14 个，字段描述如下表所示

股票数据属性

属性名称	属性说明	属性名称	属性说明
Date	日期	open	开盘价
close	收盘价	low	最低价
HIigh	最高价	volume	成交量

price_change	涨跌幅	p_change	涨跌幅差值
ma5	5 日均价	v_ma5	5 日均量
ma10	10 日均价	ma20	20 日均价
v_ma10	10 日均量	v_ma20	20 日均量

股票股价预测走势和实际股价走势图



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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：实现自定义年份分区器（20 分）

1、自定义分区器根据日期的年份对数据分区

2、程序中对象、方法关键字大小写敏感

3、编制程序时注意代码缩进凸显结构清晰

任务二：利用移动平均算法实现股票价格预测（40 分）

1、简单移动平均的计算公式为：

$$F_t = (A_{t-1} + A_{t-2} + A_{t-3} + \ldots + A_{t-n}) / n$$

Ft——对下一期的预测值

n——移动平均的时期个数

At－1，At－2,…,At－n——前 n 期的实际值

2、程序中对象、方法关键字大小写敏感

3、编制程序时注意代码缩进凸显结构清晰

4、实现移动平均的计算类 MovingAverage

5、运行参数配置时运行结果输出 HDFS 的路径正确

任务三：提交移动平均程序（20 分）

1、使用 spark-submit 提交执行 spark jar 包。

2、观察 spark jar 包执行结果

（2）实施条件

表 3-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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：实现自定义年份分区器

序号	评分内容	评分点	分值（分）
1	自定义分区器根据日期的年份对数据分区	程序编写正确	15 分
2	程序中对象、方法关键字大小写敏感	格式正确	3 分
3	编制程序时注意代码缩进凸显结构清晰	格式正确	2 分

评分项二：编写移动平均算法实现股票价格预测

序号	评分内容	评分点	分值（分）
1	程序中对象、方法关键字大小写敏感	程序的书写	2 分
2	编制程序时注意代码缩进凸显结构清晰	程序编写是否结构清晰	3 分
3	实现移动平均的计算类MovingAverage	程序编写正确	30 分
4	运行参数配置时运行结果输出HDFS的路径正确	命令参数编写正确	5 分

评分项三：提交移动平均程序

序号	评分内容	评分点	分值（分）
1	使用spark-submit提交执行spark jar包。	Spark-submit命令运行成功	15 分
2	观察spark jar包执行结果。	程序结果正确	5 分

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	实现自定义年份分区器	20分	1、自定义分区器根据日期的年份对数据分区 2、程序中对象、方法关键字大小写敏感 3、编制程序时注意代码缩进凸显结构清晰	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。
	编写移动平均算法实现股票价格预测	20分	1、程序中对象、方法关键字大小写敏感 2、编制程序时注意代码缩进凸显结构清晰 3、实现移动平均的计算类MovingAverage 4、运行参数配置时运行结果输出HDFS的路径正确	40 分	
	提交移动平均程序	20分	1、使用spark-submit提交执行spark jar包。 2、观察spark jar包执行结果。	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			

37 试题编号：3-2-7：Spark 开发-词频统计

（1）任务描述

某企业 NLP 部门需要对部分采集的文本数据进行预处理、词频统计，方便后续推荐算法模型的训练，请利用 Spark 分布式技术完成对指定文件的词频统计，并将词频统计结果的输出。

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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一: 启动集群，需要在 master, slave01,slave02 不同虚拟机上验证。

1、打开虚拟机并启动 Hadoop、Spark 集群、并取消集群的安全模式

2、在 master 上运行 jps，确认相关进程是否启动

3、在 slave01 上运行 jps，确认相关进程是否启动

4、在 slave02 上运行 jps，确认相关进程是否启动

任务二：在 idea 创建工程时配置软件包依赖和删除测试环境 test 中的测试类

1、创建工程时配置软件包依赖和删除测试环境 test 中的测试类

2、正确配置 maven 工程的 pom.xml 文件

任务三：在 idea 中完成代码编写和调试

1、程序中对象、方法关键字大小写敏感

2、编制程序时注意代码缩进凸显结构清晰

3、运行参数配置时运行结果输出 HDFS 的路径正确

（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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：开发环境搭建

序号	评分内容	评分点	分值（分）
1	启动集群并验证	启动成功	10 分
2	创建Maven工程配置软件包依赖	配置正确	5 分
3	编写pom.xml文件	pom.xml编程正确	5 分

评分项二：spark 词频统计程序

序号	评分内容	评分点	分值（分）
1	在idea中完成spark词频统计程序编写和调试	程序的书写正确	40 分
2	HDFS上输出数据是否正确	参数设置正确	20 分

Spark 大数据处理与分析模块考核评价标准

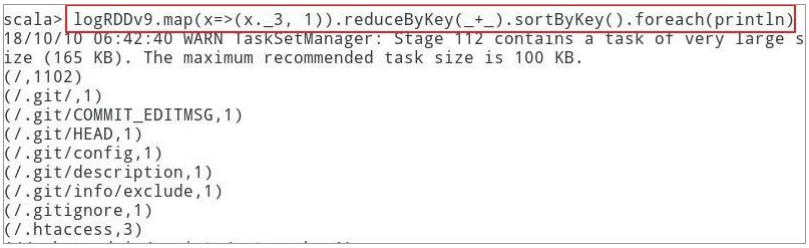
评价内容		配分	评分标准		备注
工作任务	启动集群并验证	20 分	正确启动集群并验证	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目记0分。 2、严重违反考场纪律、造成恶劣影
	创建Maven工程配置软件包依赖,编写pom.xml文件	20 分	正确创建Maven工程配置软件包依赖，编写pom.xml文件	10 分	
	开发spark词频统计程序	40 分	在idea中完成spark词频统计程序编写和调试	40 分	
	查看HDFS上输出数据是否正确	20 分	HDFS上输出数据是否正确	20 分	

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	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			

38 试题编号：3-2-8：Spark 开发-Apache 日志分析

（1）任务描述

某企业的网站部署在 Apache 服务器上，现有实际业务需求如下，请使用 Spark 完成对 Apache 格式的日志内容的分析。部分实现效果图如下：



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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一：环境准备和核心函数编写。（20 分）

1、打开终端，启动 spark-shell，，启动时指定启动模式

2、加载本地文件，使用 textFile 方法加载本地数据，生成 RDD

3、Apache 日志的一般格式:日志内容从左到右依次是：远程 IP 地址，客户端记录，浏览者记录。请求的时间，包括三项内容：，日期，时间，时区，服务器收到的请求，包括三项内容：METHOD：请求的方法，GET/POST 等。RESOURCE：请求的目标链接地址，PROTOCOL：HTTP 版本号。

4、数据预处理：获取合法的日志数据，使用正则表达式做两件事情，一个是过滤掉非法的日志，一个是解析过滤后的日志来获得需要的数据元组。

5、过滤无法解析的日志记录

6、定义解析日志的函数

任务二：统计每日 PV 数、统计独立 IP 数（20 分）

1、解析日志文件

2、统计每日 PV，使用 count 操作

3、使用 sortByKey，按照请求日期字段进行排序，并将结果保存到本地，并查看结果

4、统计独立 IP 数

任务三：统计每种不同的 HTTP 状态访问次数

1、统计每种不同的 HTTP 状态对应的访问次数。

2、降序展示。

任务四：统计不同独立 IP 的访问量、不同页面的访问量

1、统计不同独立 IP 的访问量

2、按照降序排列并展示前 10 条

3、统计不同页面的访问量。

4、由于日志中有大量的 js 文件的访问，因此我们增加一个去除列表，过滤掉属于列表中后缀名的文件的函数。

5、再对过滤后的数据，执行统计操作。

(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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：环境准备和核心函数编写

序号	评分内容	评分点	分值（分）
1	启动集群并验证	终端，启动spark-shell	10 分
2	核心函数编写	过滤无法解析的日志记录	5 分
3	解析日志函数	正确定义解析日志的函数	5 分

评分项二：统计每日 PV 数、统计独立 IP 数

序号	评分内容	评分点	分值（分）
1	解析日志文件	结果正确	5 分
2	统计每日PV，使用count操作	每日PV结果正确	5 分
3	使用sortByKey，按照请求日期字段进行排序，并将结果保存到本地，并查看结果	排序之后结果正确	5 分
4	统计独立IP数	结果正确	5 分

评分项三：统计每种不同的 HTTP 状态访问次数

序号	评分内容	评分点	分值（分）
1	统计每种不同的HTTP状态对应的访问次数	结果正确	10 分
2	降序展示	降序展示	10 分

评分项四：统计不同独立 IP 的访问量、不同页面的访问量

序号	评分内容	评分点	分值（分）
1	统计不同独立IP的访问量	结果正确	5 分
2	按照降序排列并展示前10条	展示前10条	5 分
3	统计不同页面的访问量。	统计结果正确	5 分
4	由于日志中有大量的js文件的访问，因此我们增加一个去除列表，过滤掉属于列表中后缀名的文件的函数。再对过滤后的数据，执行统计操作。	统计结果正确	5 分

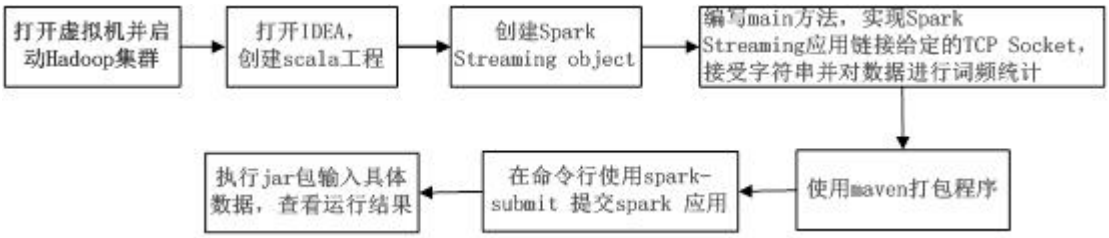
Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	环境准备和核心函数编写	20 分	1、终端，启动spark-shell 2、过滤无法解析的日志记录 3、定义解析日志的函数	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。
	统计每日PV数、统计独立IP数	20 分	1、解析日志文件 2、统计每日PV，使用count操作使用sortByKey，按照请求日期字段进行排序，并将结果保存到本地，并查看结果 3、统计独立IP数	20 分	
	统计每种不同的HTTP状态访问次数	20 分	1、统计每种不同的HTTP状态对应的访问次数。 2、降序展示。	20 分	
	统计不同独立IP的访问量、不同页面的访问量	20 分	1、统计不同独立IP的访问量 2、按照降序排列并展示前10条 3、统计不同页面的访问量。 4、由于日志中有大量的js文件的访问，因此我们增加一个去除列表，过滤掉属于列表中后缀名的文件的函数。 5、再对过滤后的数据，执行统计操作。	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			

39 试题编号：3-2-9：Spark 开发-SparkStreaming 实时网络处理数据

（1）任务描述

某企业需要实时计算网络文本数据进行预处理、词频统计，方便后续推荐算法模型的训练，请利用 SparkStreaming 分布式技术完成对网络数据的实时词频统计，并将词频统计结果的输出。实验流程步骤图如下：



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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一: 启动 Spark 集群,需要在 master, slave01,slave02 不同虚拟机上验证。（10 分）

1、打开虚拟机并启动 Hadoop 集群、并取消集群的安全模式

2、在 master 上运行 jps，确认 NameNode, SecondaryNameNode, ResourceManager 进程启动

3、在 slave01 上运行 jps，确认 DataNode, NodeManager 进程启动

4、在 slave02 上运行 jps，确认 DataNode, NodeManager 进程启动

任务二：在 idea 创建 Maven 工程时配置 SparkStreaming 软件包依赖。（10 分）

1、创建工程时配置软件包依赖和删除测试环境 test 中的测试类

2、正确配置 maven 工程的 pom.xml 文件

任务三:在 idea 中完成代码编写和调试,使用 Maven 完成打包并提交执行 spark jar 包。（40 分）

1、程序中对象、方法关键字大小写敏感

2、编制程序时注意代码缩进凸显结构清晰

3、运行参数配置时运行结果输出 HDFS 的路径正确

任务四：启动 nc -lk 9999 网络工具，并发送数据。（20 分）

1、nc -lk 9999 网络工具发送数据

2、观察 spark jar 包执行结果

（2）实施条件

表 3-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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版
4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：环境准备和核心函数编写

序号	评分内容	评分点	分值（分）
1	启动集群并验证	正确启动	10 分
2	创建Maven工程配置软件包依赖，编写pom.xml文件	正确配置	10 分

评分项二：开发 spark Streaming 程序

序号	评分内容	评分点	分值（分）
1	编写spark Streaming程序	1、在idea中完成代码编写和调试 2、使用maven工程完成打包 3、使用spark-submit提交执行spark jar包。	40 分
2	启动nc -lk 9999网络工具	启动是否成功	10 分
3	nc -lk 9999网络工具，发送数据	1、nc -lk 9999网络工具发送数据 2、观察spark jar包执行结果	10 分

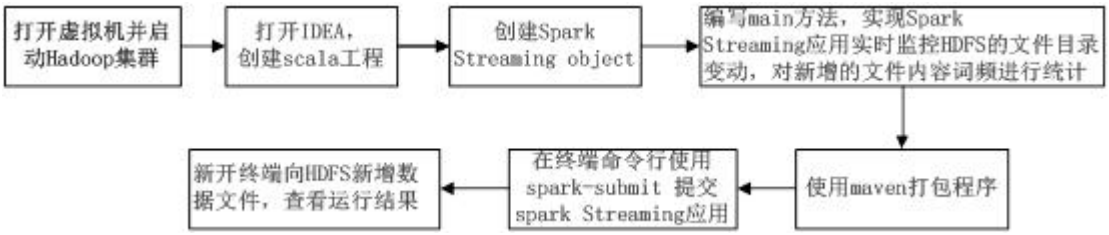
Spark 大数据处理与分析模块考核评价标准

评价内容		配分	评分标准		备注
工作任务	启动集群并验证	20分	正确启动集群并验证	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。
	创建Maven工程配置软件包依赖，编写pom.xml文件	20分	正确创建Maven工程配置软件包依赖，编写pom.xml文件	10 分	
	开发spark Streaming程序	20分	1、在idea中完成代码编写和调试 2、使用maven工程完成打包 3、使用spark-submit提交执行spark jar包。	40 分	
	启动nc -lk 9999网络工具，并发送数据	20分	1、nc -lk 9999网络工具发送数据 2、观察spark jar包执行结果	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			

40 试题编号：3-2-10：Spark 开发-SparkStreaming 实时 HDFS 处理数据

（1）任务描述

某企业需要实时计算 flume 等软件收集的网络日记文本数据，词频统计，方便后续推荐算法模型的训练，请利用 SparkStreaming 分布式技术完成对网络数据的实时词频统计，并将词频统计结果的输出。实验流程步骤图如下：



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：湖南信息职业技术学院 Spark 大数据处理与分析 01 张三答案.docx。

任务一: 启动 Spark 集群,需要在 master, slave01,slave02 不同虚拟机上验证。（10 分）

1、打开虚拟机并启动 Hadoop 集群、并取消集群的安全模式

2、在 master 上运行 jps，确认 NameNode, SecondaryNameNode, ResourceManager 进程启动

3、在 slave01 上运行 jps，确认 DataNode, NodeManager 进程启动

4、在 slave02 上运行 jps，确认 DataNode, NodeManager 进程启动

任务二：创建 Maven 工程配置软件包依赖，编写 pom.xml 文件（10 分）

1、在 idea 中 maven 工程并配置好 maven 仓库地址

2、正确配置 maven 工程的 pom.xml 依赖

任务三：使用 IDEA 开发 spark Streaming 程序，使用 Maven 完成打包并提交执行 spark jar 包。（40 分）

1、在 idea 中完成代码编写和调试

2、使用 maven 工程完成打包

3、使用 spark-submit 提交执行 spark jar 包。

任务四：上传数据到 HDFS，并观察实验结果数据。（20 分）

1、上传数据到 hdfs。

2、观察 spark jar 包执行结果

（2）实施条件

表 3-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	VMware Workstation	12.0 或以上	12.0 后的系统必须安装在 64 位操作系统中
3	办公软件	Microsoft Office 2007	可以高于 2007 版

4	远程登录软件	SecureCRT 或 Xshell	用于远程连接 centos
5	Linux 安装光盘镜像	CENTOS 7.2 及以上	用于在虚拟机中安装操作系统

（3）考核时量

考核时间为 150 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：

评分项一：环境准备和核心函数编写

序号	评分内容	评分点	分值（分）
1	启动集群并验证	正确启动	10 分
2	创建Maven工程配置软件包依赖，编写pom.xml文件	正确配置	10 分

评分项二：开发 spark Streaming 程序

序号	评分内容	评分点	分值（分）
1	开发spark Streaming程序	1、在idea中完成代码编写和调试 2、使用maven工程完成打包 3、使用spark-submit提交执行spark jar包。	40 分
2	上传数据到HDFS，并观察实验结果数据	1、上传数据到hdfs。 2、观察spark jar包执行结果	20 分

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	启动Spark集群并验证	10 分	正确启动Spark集群并验证	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
	创建Maven工程配置软件包依赖，编写pom.xml文件	10 分	正确创建Maven工程配置软件包依赖，编写pom.xml文件	10 分	
	开发spark Streaming程序	40 分	1、在idea中完成代码编写和调试 2、使用maven工程完成打包 3、使用spark-submit提交执行	40 分	

			spark jar包。		记0分。
	上传数据到HDFS，并观察实验结果数据	20分	1、上传数据到hdfs。 2、观察spark jar包执行结果	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			

总计	100 分

模块四 数据分析与可视化

项目 1：基于 matplotlib 的数据分析和可视化

41.试题编号：4-1-1，单日票房数据分析和可视化

（1）任务描述

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

任务描述

1.导入数据分析和可视化需用的相关模块，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。

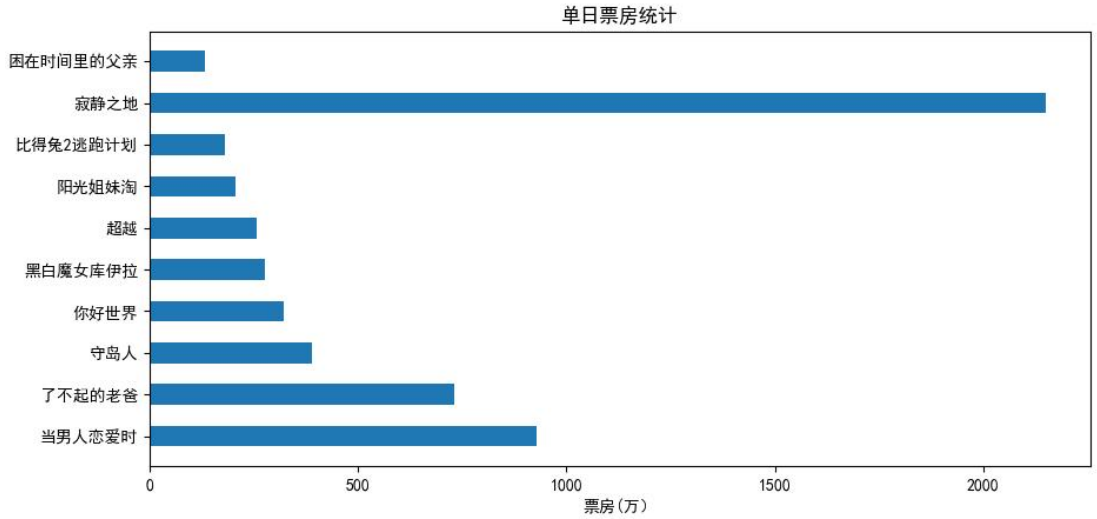


图 4-1-2　单日票房统计

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,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 120 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 matplotlib 库正确 5 分 导入 pandas 库正确 5 分	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
	设置 rc 参数	5 分	图表显示中文设置正确 5 分	5 分	
	读文件及筛选数据	10 分	读取文件内容正确 5 分 筛选数据操作正确 5 分	10 分	
	绘制图形	15 分	函数名称正确 5 分 函数参数传递正确 10 分	15 分	
	设置标题	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置横坐标	10 分	函数名称正确 5 分	10 分	

			函数参数传递正确 5 分		记0分。
	设置纵坐标	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	保存和显示图表	10 分	保存图表函数调用正确 5 分 显示图表函数调用正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	
总计		100 分			

42.试题编号：4-1-2，单周票房数据分析和可视化

（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

任务要求

1.导入数据分析和可视化需用的相关模块，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。

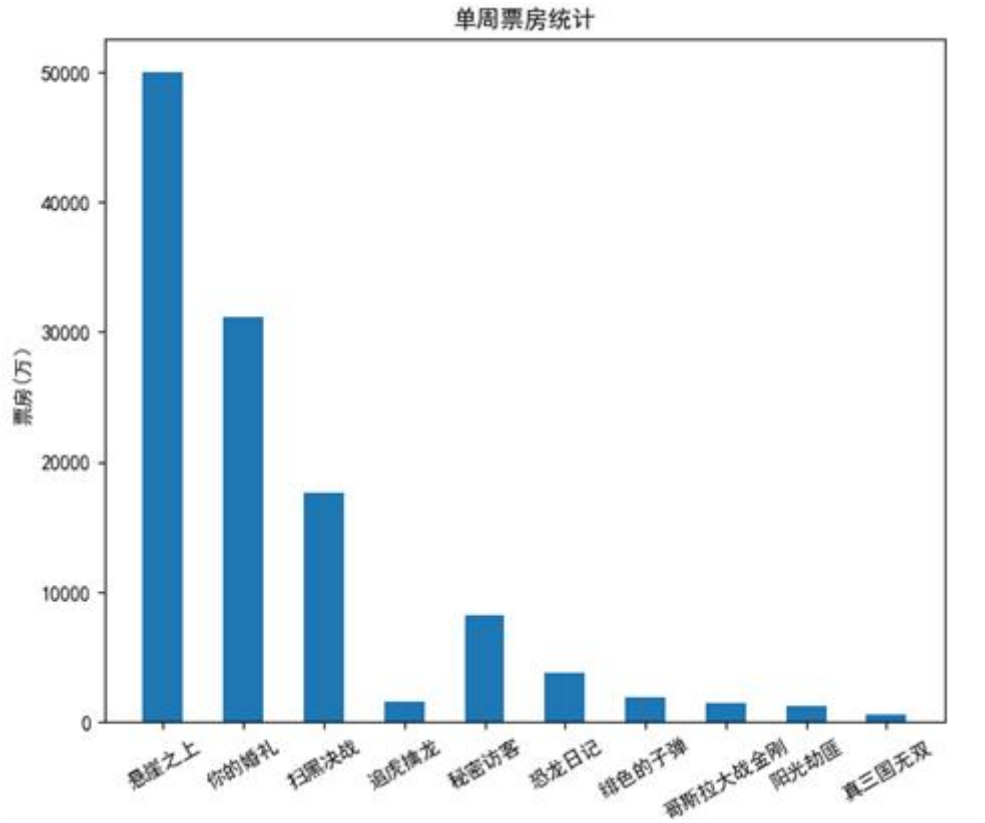


图 4-2-2 单周票房统计

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，

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。

（2）实施条件

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、numpy,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 120 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 matplotlib 库正确 5 分 导入 pandas 库正确 5 分	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	图表显示中文设置正确 5 分	5 分	
	读文件及筛选数据	10 分	读取文件内容正确 5 分 筛选数据操作正确 5 分	10 分	
	绘制图形	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置标题	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置横坐标	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置纵坐标	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置图例	5 分	函数名称正确 3 分 函数参数传递正确 2 分	5 分	
	保存和显示图表	10 分	保存图表函数调用正确 5 分 显示图表函数调用正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	
总计		100 分			

43.试题编号：4-1-3，单月票房数据分析和可视化

（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

任务要求

1.导入数据分析和可视化需用的相关模块，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。

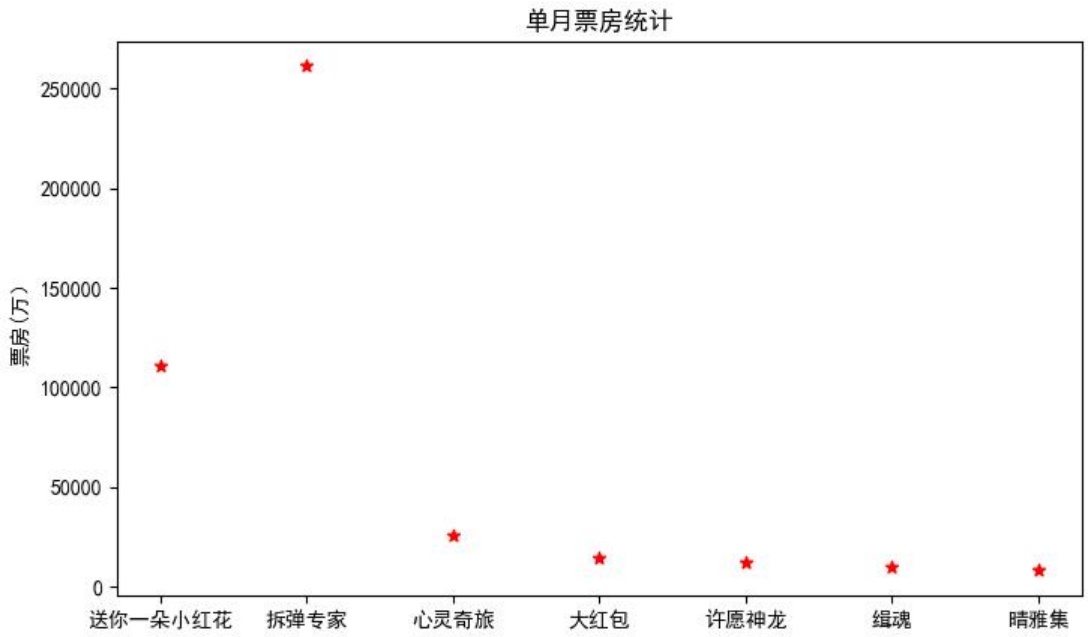


图 4-3-2 单月票房统计

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,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		测评专家满足任一条件

	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）。	

（3）考核时量

考核时间为 120 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 matplotlib 库正确 5 分 导入 pandas 库正确 5 分	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	图表显示中文设置正确 5 分	5 分	
	读文件及筛选数据	10 分	读取文件内容正确 5 分 筛选数据操作正确 5 分	10 分	
	绘制图形	15 分	函数名称正确 5 分 函数参数传递正确 10 分	15 分	
	设置标题	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置横坐标	10 分	函数名称正确 5 分 函数参数传递正确 10 分	10 分	
	设置纵坐标	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	保存和显示图表	10 分	保存图表函数调用正确 5 分 显示图表函数调用正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分	0-10 分	

			注释规范3分		
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10分	
总计		100 分			

44.试题编号：4-1-4，档期总票房数据分析和可视化

（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,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

1.导入数据分析和可视化需用的相关模块，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。

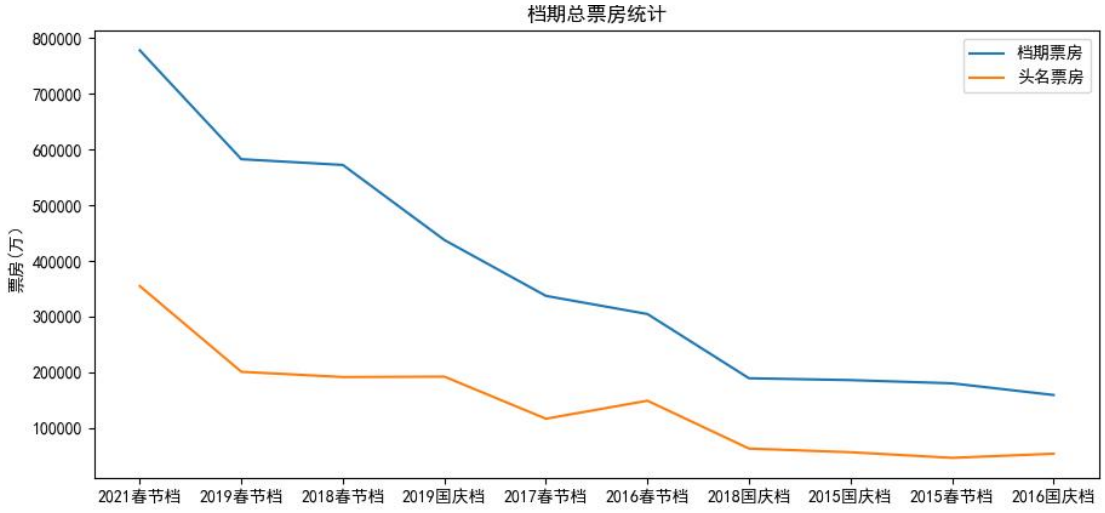


图 4-4-2 档期总票房

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,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 120 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 matplotlib 库正确 5 分 导入 pandas 库正确 5 分	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	图表显示中文设置正确 5 分	5 分	
	读文件及筛选数据	10 分	读取文件内容正确 5 分 筛选数据操作正确 5 分	10 分	
	绘制图形	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置标题	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置横坐标	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置纵坐标	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置图例	5 分	函数名称正确 3 分 函数参数传递正确 2 分	5 分	

	保存和显示图表	10 分	保存图表函数调用正确 5 分 显示图表函数调用正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	
总计		100 分			

45.试题编号：4-1-5，内地总票房排名数据分析和可视化

（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

任务要求

1.导入数据分析和可视化需用的相关模块，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。

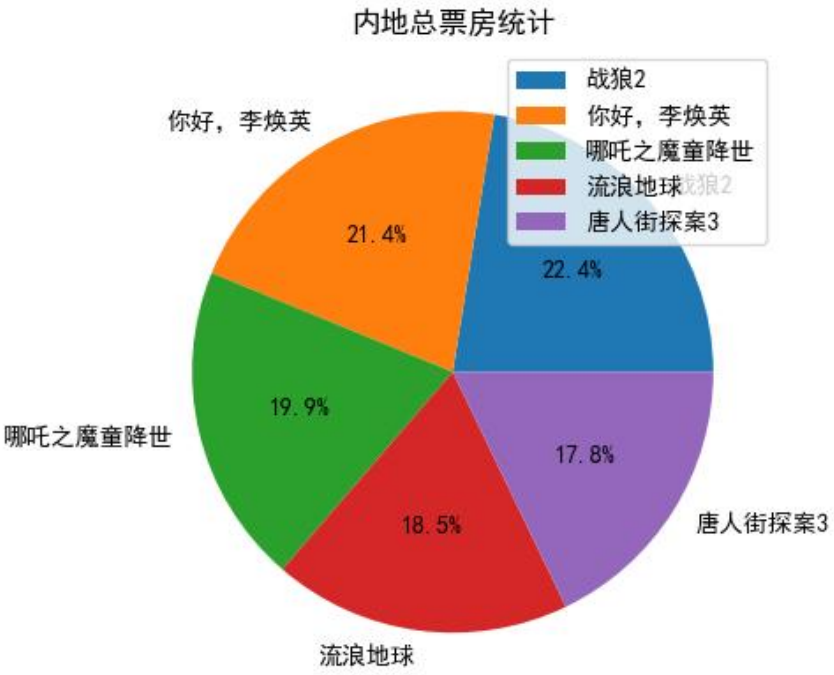


图 4-5-2　内地总票房

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,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 120 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 matplotlib 库正确 5 分 导入 pandas 库正确 5 分	10 分	1、考试舞弊、抄袭、没有按要求填写相关信息，本项目
	设置 rc 参数	5 分	图表显示中文设置正确 5 分	5 分	
	读文件及筛选数据	10 分	读取文件内容正确 5 分 筛选数据操作正确 5 分	10 分	
	绘制图形	25 分	函数名称正确 10 分	25 分	

			函数参数传递正确 15 分		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	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	设置图例	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
	保存和显示图表	10 分	保存图表函数调用正确 5 分 显示图表函数调用正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	
总计		100 分			

项目 2：基于 pyecharts 的数据可视化

46.试题编号：4-2-1，空气质量指数 AQI 数据可视化

（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	SO2	CO	NO2	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；

SO2：二氧化硫平均浓度值；

CO：一氧化碳硫平均浓度值；

NO2：二氧化氮硫平均浓度值；

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 4-6-1 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”。

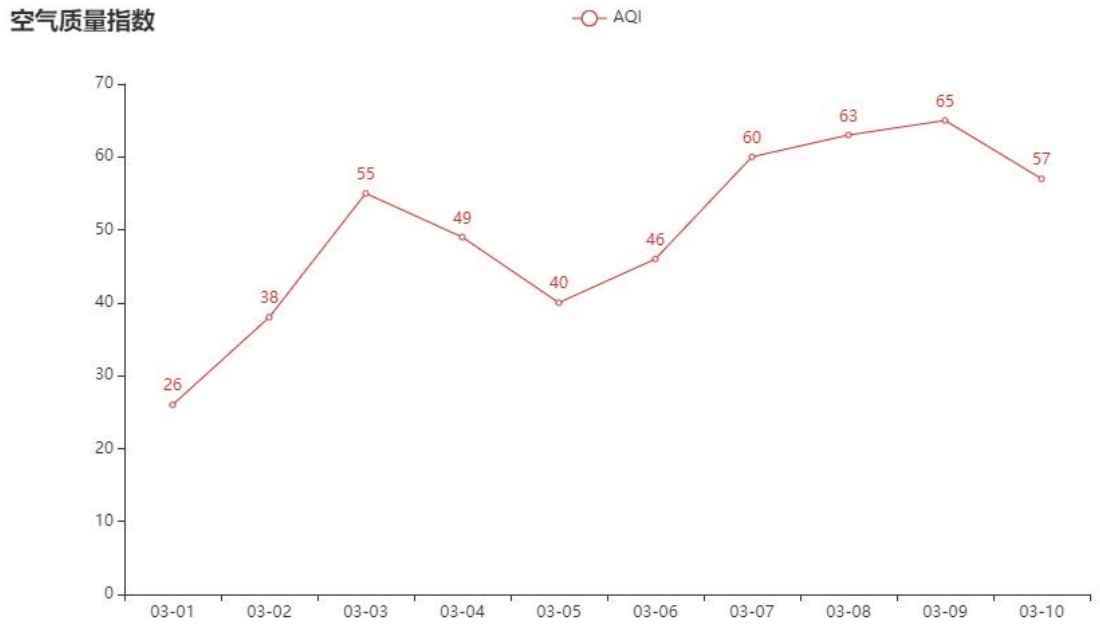


图 4-6-1 空气质量指数

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,pandas、pyecharts1.9.0、pyecharts_snapshot）、火狐浏览器或谷歌浏览器	
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）。		

（3）考核时量

考核时间为 120 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	图表类导入正确 5 分 option 模块导入正确 5 分	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
	保存数据	30 分	x 轴数据选择正确 5 分 保存 x 轴数据操作正确 10 分 y 轴数据选择正确 5 分 保存 y 轴数据操作正确 10 分	10 分	
	绘制图形	20 分	函数名称正确 10 分 函数参数传递正确 10 分	20 分	
	设置标题	10 分	参数名称正确 5 分 参数赋值正确 5 分	10 分	
	保存图表	10 分	函数名称正确 5 分	10 分	

			函数参数传递正确 5 分		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5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	
总计		100 分			

47.试题编号：4-2-2，空气质量 NO2 数据可视化

（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	SO2	CO	NO2	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；

SO2：二氧化硫平均浓度值；

CO：一氧化碳硫平均浓度值；

NO2：二氧化氮硫平均浓度值；

03_8h：臭氧 8 小时平均浓度值

任务要求

1.导入绘图需用的相关模块，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 4-7-1 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”。

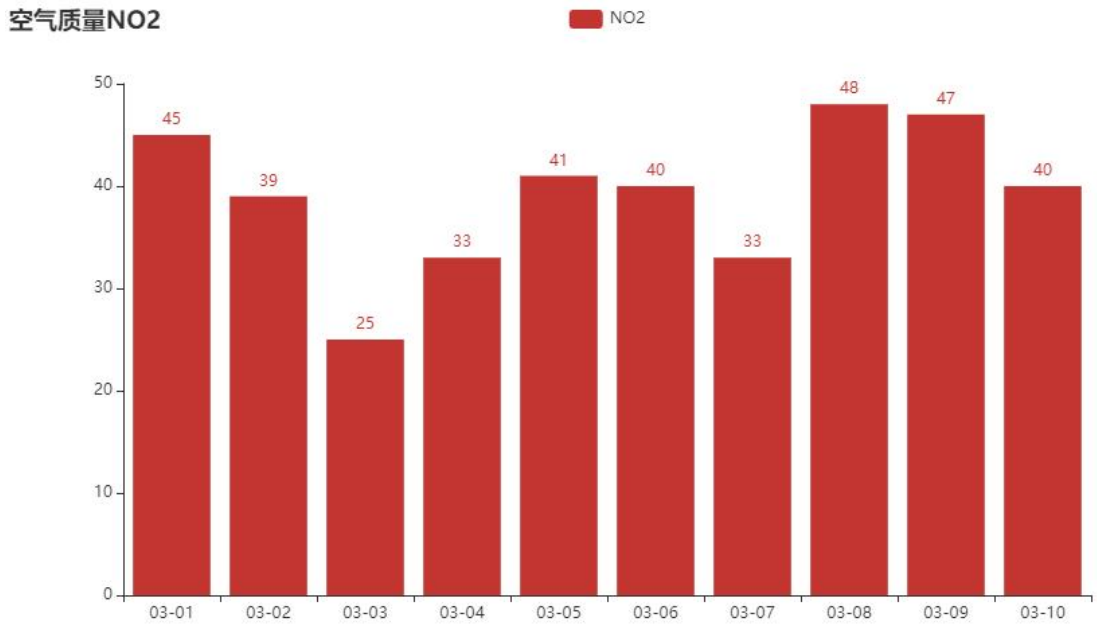


图 4-7-1 空气质量 NO2

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，

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。

（2）实施条件

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、numpy,pandas、pyecharts1.9.0、pyecharts_snapshot）、火狐浏览器或谷歌浏览器	
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）。		

（3）考核时量

考核时间为 120 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	图表类导入正确 5 分 option 模块导入正确 5 分	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。
	保存数据	30 分	x 轴数据选择正确 5 分 保存 x 轴数据操作正确 10 分 y 轴数据选择正确 5 分 保存 y 轴数据操作正确 10 分	10 分	
	绘制图形	20 分	函数名称正确 10 分 函数参数传递正确 10 分	20 分	
	设置标题	10 分	参数名称正确 5 分 参数赋值正确 5 分	10 分	
	保存图表	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	
总计		100 分			

48.试题编号：4-2-3，空气质量 PM10 数据可视化

（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 4-8-1 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”。

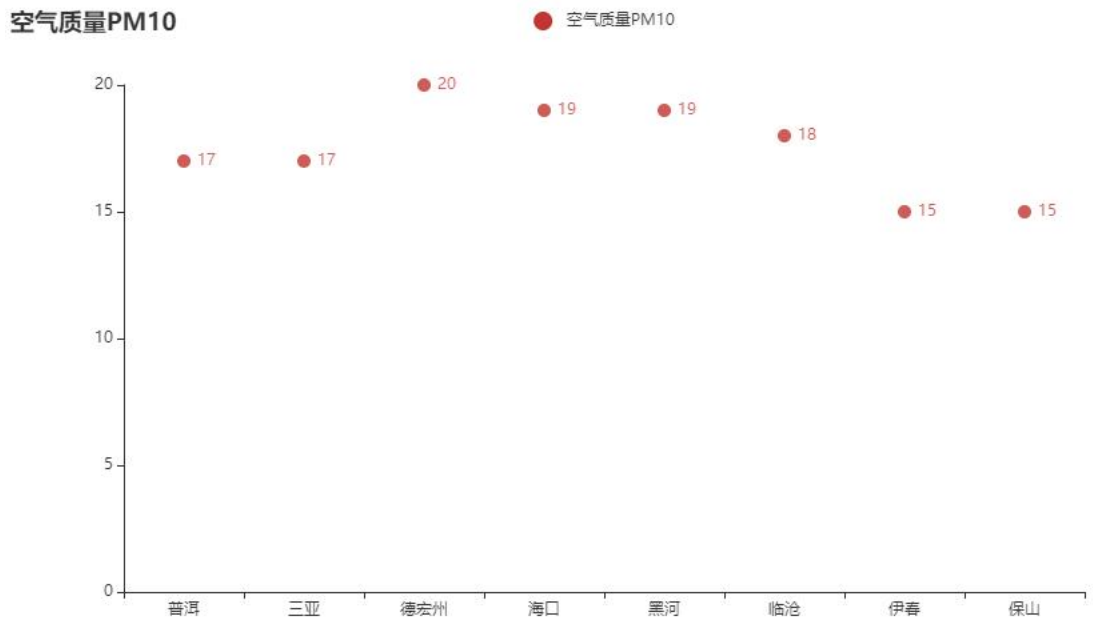


图 4-8-1 空气质量 PM10

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,pandas、pyecharts1.9.0、pyecharts_snapshot）、火狐浏览器或谷歌浏览器	
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）。		

（3）考核时量

考核时间为 120 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	图表类导入正确 5 分 option 模块导入正确 5 分	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。
	保存数据	30 分	x 轴数据选择正确 5 分 保存 x 轴数据操作正确 10 分 y 轴数据选择正确 5 分 保存 y 轴数据操作正确 10 分	10 分	
	绘制图形	20 分	函数名称正确 10 分 函数参数传递正确 10 分	20 分	
	设置标题	10 分	参数名称正确 5 分 参数赋值正确 5 分	10 分	
	保存图表	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	
总计		100 分			

49.试题编号：4-2-4，空气质量 PM2.5 数据可视化

（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；

1.导入绘图需用的相关模块，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.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”。

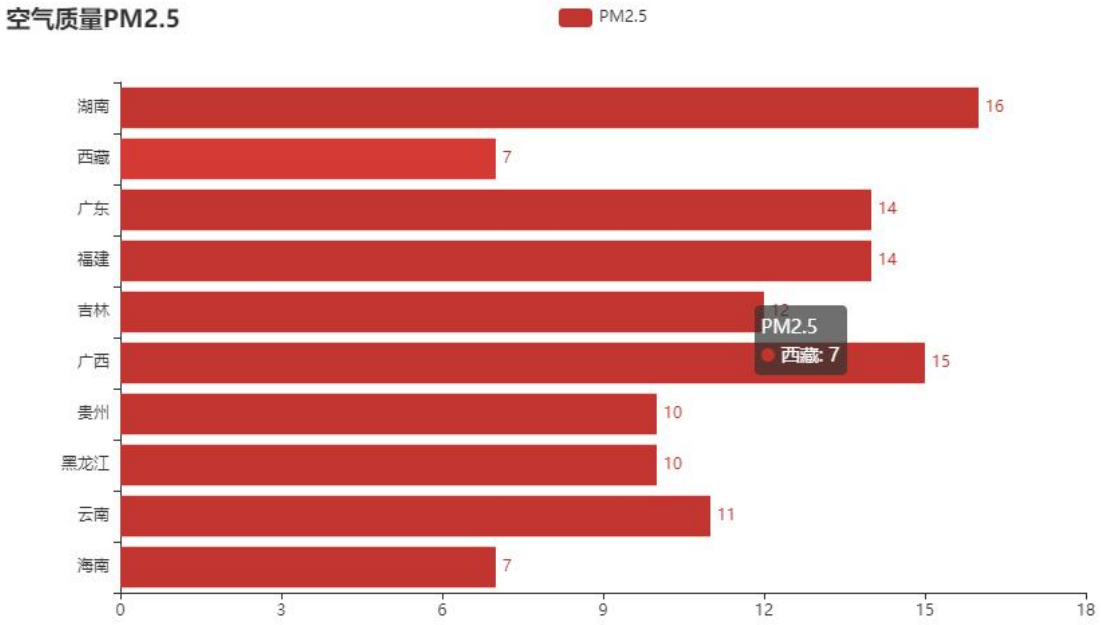


图 4-9-1 空气质量 PM2.5

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,pandas、pyecharts1.9.0、pyecharts_snapshot）、火狐浏览器或谷歌浏览器	
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		测评专家满足任一条件

	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）。	

（3）考核时量

考核时间为 120 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	图表类导入正确 5 分 option 模块导入正确 5 分	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。
	保存数据	20 分	x 轴数据选择正确 5 分 保存 x 轴数据操作正确 5 分 y 轴数据选择正确 5 分 保存 y 轴数据操作正确 5 分	10 分	
	绘制图形	20 分	函数名称正确 10 分 函数参数传递正确 10 分	20 分	
	设置标题	10 分	参数名称正确 5 分 参数赋值正确 5 分	10 分	
	设置数据标签显示位置	10 分	参数名称正确 5 分 参数赋值正确 5 分		
	保存图表	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分 缩进统一，方便阅读1分 注释规范3分	0-10 分	
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10 分	

总计	100 分

50.试题编号：4-2-5，空气质量指数 AQI 和 PM2.5 数据可视化

（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；

任务要求

1.导入绘图需用的相关模块，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 4-10-1 中的省份列。

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”。

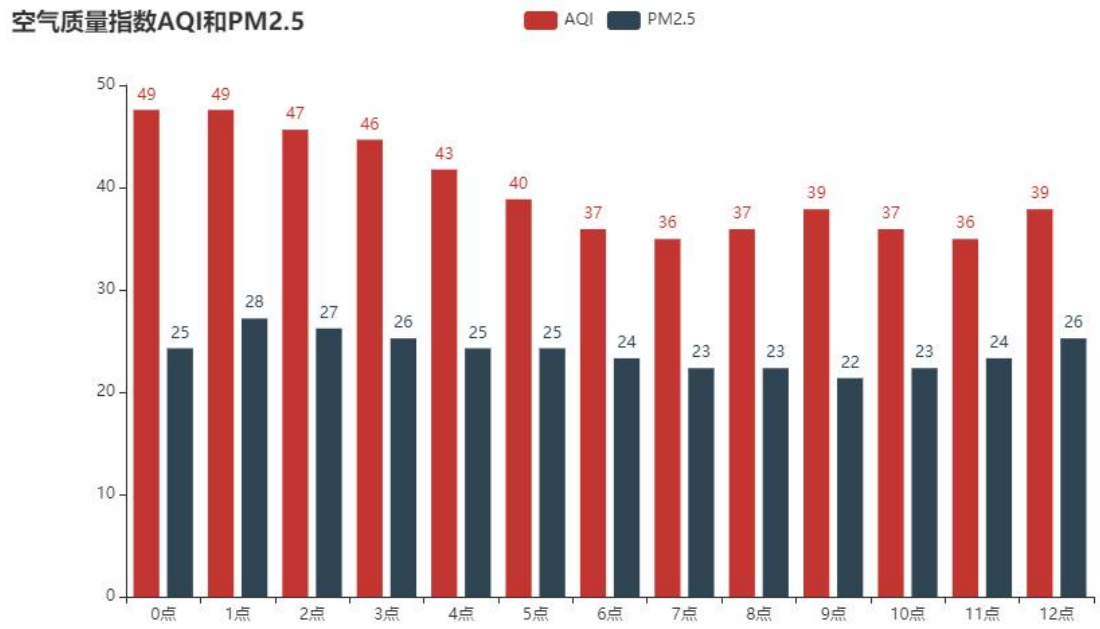


图 4-10-1 空气质量指数 AQI 和 PM2.5

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：湖南信息职业技术学院 01 张三。

2）“技能抽查提交资料”文件夹内保存代码源文件及引用的相关素材文件，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。

（2）实施条件

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、numpy,pandas、pyecharts1.9.0、pyecharts_snapshot）、火狐浏览器或谷歌浏览器	
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）。		

（3）考核时量

考核时间为 120 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	图表类导入正确 5 分 option 模块导入正确 5 分	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。
	保存数据	30 分	x 轴数据选择正确 5 分 保存 x 轴数据操作正确 10 分 y 轴数据选择正确 5 分 保存 y 轴数据操作正确 10 分	10 分	
	绘制图形	20 分	函数名称正确 10 分 函数参数传递正确 10 分	20 分	
	设置标题	10 分	参数名称正确 5 分 参数赋值正确 5 分	10 分	
	保存图表	10 分	函数名称正确 5 分 函数参数传递正确 5 分	10 分	
职业素养	专业素养	10 分	代码符合代码开发规范5分 命名规范，能做到见名知意1分	0-10 分	

			缩进统一，方便阅读1分 注释规范3分		
	道德规范	10 分	着装干净、整洁5分 举止文明，遵守考场纪律，按顺序进出考场5分	0-10分	
总计		100 分			