ZZ052-大数据应用与服务赛项试题 07

一、背景描述

当今时代，数据正在迅速膨胀并变大，一天之中，互联网产生的全部内容可以达到 EB 级别，能够轻松刻满 1.68 亿张光盘。在商业、经济及其它领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉。那么，要怎样基于大数据做出正确的决策呢？大数据首先需要解决的问题就是数据存储的问题，由于数据量非常之大，想通过传统单一的节点的存储显得力不从心，搭建分布式的文件存储系统成为了一个完美的解决方案。解决了数据存储的问题，我们需要从数据中提取有用信息，通过数据分析手段让数据发挥出真正的价值。但往往采集的原始数据中包含了一些无用数据以及噪声数据，如果直接基于这些脏数据进行分析，往往会让分析结果产生偏差甚至错误，从而造成决策上的失准。因此，我们有必要对这些原始数据进行清洗，以保证其数据准确性、完整性和可用性，提高数据的质量。在解决脏数据的困扰后，我们需要采取各种数据分析手段，提取数据中的价值，得到可靠的结果，并以图表等直观的方式将分析结果进行展现。然后从业务层面对分析结果进行分析和解释，从而指引我们做出正确的决策，真正获取“数据财富”。

气候变化正在迅速地改变地球。随着全球气温不断升高、

1

海平面上升、极端天气事件频繁发生，人们对于地球的未来更加担忧。为了更好地了解气候变化的趋势、预测未来天气趋势，指引相关部门尽早做出举措以应对气候变化，保护人类赖以生存的家园，你的团队将运用大数据技术对天气数据进行分析及决策。搭建大数据平台集群环境以应对海量天气数据的存储，结合数据库的毫秒级的响应，为天气决策系统提供数据存储及查询保障。通过数据清洗技术，去除数据中的噪音，提高数据质量。通过数据标注技术，结合业务认知，对数据进行分类标注，为后续通过人工智能算法模型决策奠定基础。通过各种数据分析技术，让看似杂乱无章的数据，变得灵动，找出天气变化的内在规律。通过数据可视化技术，让数据分析结果及天气变化规律以一种最为直观的方式呈现。最后从业务层面对天气数据分析结果进行分析及解释，使气象学家更好的了解气候变化，并做出精准决策应对气候问题。你们作为该大数据小组的技术人员，请按照下面任务完成本次工作。

二、模块一：平台搭建与运维

（一）任务一：大数据平台搭建

1.子任务一：Hadoop 完全分布式安装配置

本任务需要使用 root 用户完成相关配置，安装 Hadoop 需要配置前置环境。命令中要求使用绝对路径，具体要求如下:

（1）从 Master 中的/opt/software 目录下将文件

hadoop-3.1.3.tar.gz、jdk-8u191-linux-x64.tar.gz 安装包解压到/opt/module 路径中(若路径不存在，则需新建)，将 JDK 解压命令复制并粘贴至客户端桌面【M1-T1-SUBT1-提交结果 1.docx】中对应的任务序号下;

（2）修改 Master 中/etc/profile 文件，设置 JDK 环境变量并使其生效，配置完毕后在 Master 节点分别执行“java -version”和“javac”命令，将命令行执行结果分别截图并粘贴至客户端桌面【M1-T1-SUBT1-提交结果 2.docx】中对应的任务序号下;

（3）请完成 host 相关配置，将三个节点分别命名为 master、slave1、slave2，并做免密登录，用 scp 命令并使用绝对路径从 Master 复制 JDK 解压后的安装文件到 slave1、slave2 节点（若路径不存在，则需新建），并配置 slave1、slave2 相关环境变量，将全部 scp 复制 JDK 的命令复制并粘贴至客户端桌面【M1-T1-SUBT1-提交结果 3.docx】中对应的任务序号下;

（4）在 Master 将 Hadoop 解压到/opt/module(若路径不存在，则需新建)目录下，并将解压包分发至 slave1、slave2 中，其中 master、slave1、slave2 节点均作为 datanode，配置好相关环境，初始化 Hadoop 环境 namenode，将初始化命令及初始化结果截图（截取初始化结果日志最后 20 行即可）粘贴至客户端桌面【M1-T1-SUBT1-提交结果 4.docx】中对应的任务序号下;

3

（5）启动 Hadoop 集群（包括 hdfs 和 yarn），使用 jps 命令查看 Master 节点与 slave1 节点的 Java 进程，将 jps 命令与结果截图粘贴至客户端桌面【M1-T1-SUBT1-提交结果 5.docx】中对应的任务序号下。

2.子任务二：Flume 安装配置

本任务需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体要求如下：

（1）从 Master 中的/opt/software 目录下将文件 apache-flume-1.9.0-bin.tar.gz 解压到/opt/module 目录下，将解压命令复制并粘贴至客户端桌面【M1-T1-SUBT2-提交结果 1.docx】中对应的任务序号下；

（2）完善相关配置设置，配置 Flume 环境变量，并使环境变量生效，执行命令 flume-ng version 并将命令与结果截图粘贴至客户端桌面【M1-T1-SUBT2-提交结果 2.docx】中对应的任务序号下；

（3）启动 Flume 传输 Hadoop 日志（namenode 或 datanode 日志），查看 HDFS 中/tmp/flume 目录下生成的内容，将查看命令及结果（至少 5 条结果）截图粘贴至客户端桌面【M1-T1-SUBT2-提交结果 3.docx】中对应的任务序号下。

3.子任务三：Flink on Yarn 安装配置

本任务需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体要求如下：

（1）从 Master 中的/opt/software 目录下将文件

flink-1.14.0-bin-scala_2.12.tgz 解压到路径/opt/module 中(若路径不存在，则需新建)，将完整解压命令复制粘贴至客户端桌面【M1-T1-SUBT3-提交结果 1.docx】中对应的任务序号下;

（2）修改容器中/etc/profile 文件，设置 Flink 环境变量并使环境变量生效。在容器中/opt 目录下运行命令 flink --version，将命令与结果截图粘贴至客户端桌面【M1-T1-SUBT3-提交结果 2.docx】中对应的任务序号下;

（3）开启 Hadoop 集群，在 yarn 上以 per job 模式（即 Job 分离模式，不采用 Session 模式）运行 $FLINK_HOME/examples/batch/WordCount.jar，将运行结果最后 10 行截图粘贴至客户端桌面【M1-T1-SUBT3-提交结果 3.docx】中对应的任务序号下。

（二）任务二: 数据库配置维护

1.子任务一: 创建数据库及相关数据表

在 MySQL 数据库中创建“test”数据库，并在“test”数据库中分别创建“stu”、“course”及“score”共 3 个数据表。各个数据表的表字段格式如下:

表 1“stu”的表字段结构

字段	类型	备注
学号	varchar	主键
姓名	varchar	
性别	varchar	
专业	varchar	
班级	varchar	
学院	varchar	

表 2“course”的表字段结构

字段	类型	备注
课程号	varchar	主键
课程名称	varchar	
开设学院	varchar	
学分	int	

5

表 3“score”的表字段结构

字段	类型	备注
学号	varchar	联合主键
课程号	varchar	
成绩	double	

将创建“test”数据库、“stu”、“course”及“score”的建表结果图分别截图复制粘贴至客户端桌面【M1-T2-SUBT1-提交结果 1.docx】中对应的任务序号下。

2.子任务二：添加数据记录

分别为“stu”、“course”及“score”数据表添加数据记录。各个数据表所需要添加的数据记录如下:

表 4“stu”数据表的数据记录

学号	姓名	性别	专业	班级	学院
2020010101	黄洋华	男	计算机	20 计算机 1 班	电子
2021020201	张明洋	男	物联网	21 物联网 2 班	电子
2022030105	章小明	女	市场营销	22 市营 1 班	经管
2021040306	宝明文	男	机器人	21 机器人 1 班	智能
2022030212	曲飞飞	女	市场营销	22 市营 1 班	经管
2022050219	陈大华	男	电气自动化	22 电气 1 班	智能
2021010423	徐宝文	男	计算机	21 计算机 1 班	电子
2022080229	赵宝宝	女	会计	22 会计 1 班	经管

表 5“course”数据表的数据记录

课程号	课程名称	开设学院	学分
KCDZ01	C 语言程序设计	电子学院	3
KCJG01	会计大数据分析	经管学院	3
KJZN01	自动控制应用	智能学院	3
KCDZ02	人工智能概论	电子学院	2
KJJG02	市场营销实践	经管学院	2

表 6“score”数据表的数据记录

学号	课程号	成绩
2020010101	KCJG01	78
2020010101	KJZN01	89

6

2020010101	KCDZ02	65
2021020201	KCJG01	76
2021020201	KJZN01	79
2021020201	KCDZ02	89
2021020201	KJJG02	45
2022030105	KCDZ01	85
2022030105	KJZN01	68
2022030105	KCDZ02	48
2021040306	KCDZ02	92
2021040306	KCDZ01	75
2021040306	KCJG01	69
2022030212	KCDZ01	77
2022030212	KCDZ02	81
2022030212	KJZN01	63
2022050219	KCDZ01	86
2022050219	KCDZ02	72
2022050219	KJJG02	77
2021010423	KCDZ01	91
2021010423	KCJG01	64
2021010423	KJZN01	83
2022080229	KCDZ01	86
2022080229	KCJG01	70
2022080229	KCDZ02	50
2022080229	KJJG02	62

3.子任务三：数据表查询

（1）将学院名称为“电子”的所有学生其“学号”、“姓名”、“课程名称”及“成绩”显示出来;

（2）将选修了课程号为“KCJG01”和“KCDZ02”的所有学生其“姓名”、“课程名称”及“成绩”显示出来;

（3）将姓名末尾带有“华”字的学生其“姓名”、“课程名称”及“成绩”显示出来。

将上述SQL查询语句及查询结果图分别截图复制粘贴至客户端桌面【M1-T2-SUBT3-提交结果1.docx】中对应的任务序号下。

7

三、模块二：数据获取与处理

（一）任务一：数据获取与清洗

1.子任务一：数据获取

打开 ZZ052-7-M2-T1-SUBT1 文件夹，文件夹中包含 train.csv 文件。train.csv 文件是一份银行客户的数据，包括：年龄、职业、婚姻、教育情况、信用卡是否有违约、是否有房贷、是否有贷款、联系方式、上一次联系的月份、上一次联系的星期几、上一次联系的时长（秒）、活动期间联系客户的次数、上一次与客户联系后的间隔天数、在本次营销活动前，与客户联系的次数、之前营销活动的结果、雇员人数（季度指标）、就业变动率（季度指标）、消费者价格指数（月度指标）、消费者信心指数（月度指标）、银行同业拆借率 3 个月利率（每日指标）、客户是否进行购买。使用 pandas 读取 train.csv 并将读取的结果打印在 IDE 终端上，读取代码的截图复制粘贴至客户端桌面【M2-T1-SUBT1-提交结果 1.docx】中对应的任务序号下。

2.子任务二：数据处理

现已从相关网站及平台获取到原始数据集，为保障用户隐私和行业敏感信息，已进行数据脱敏。数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。在涉及客户安全数据或者一些商业性敏感数据的情况、不违反系统规则条件下，对真实数据进行改造并提供测试使用，如身份证号、手机号等个人信息都需要进行

8

数据脱敏。

打开 ZZ052-7-M2-T1-SUBT1 文件夹，文件夹中包含 train.csv 文件。你的小组需要通过编写代码或脚本完成对相关数据文件中数据的清洗和整理。请分析相关数据集，根据题目规定要求实现数据处理，具体要求如下:

（1）查看 train.csv 中数据总数、标准差、均值、最小值、四分之一分位数、二分之一分位数、四分之三分位数和最大值;

将上述代码截图复制粘贴至客户端桌面【M2-T1-SUBT2-提交结果 1.docx】中对应的任务序号下。

（2）缺失值处理:

①对于 job 列数据，采用 ‘admin.’ 填充缺失值;

②对于 marital 列数据，如果年龄（age）小于 30 采用′single′，如果大于 50 采用′divorced′代替，其他采用′marital′代替;

③将教育类型 basic.9y,basic.6y,basic.4y unknown 均变为 Basic;

④对于 housing 列数据，如果信用卡是有违约，即 default 为 yes，则用 yes 代替，否则用 no 代替;

⑤对于 loan 列数据,如果有房贷,即 housing 为 yes，则用 yes 代替，否则用 no 代替。

所有缺失值处理完后，存入 train_c1.csv 中。

9

将上述①-⑤任务的代码截图及结果截图复制粘贴至客户端桌面【M2-T1-SUBT2-提交结果2.docx】中对应的任务序号下。

（3）查看 train.csv 中的数字特征，对数字特征进行描述性统计，并采用四分位数法进行数据清洗以减少噪声数据的影响，然后存入 train_c2.csv 中;

（4）对 train.csv 中的非数字特征进行 LabelEncoder 编码并存入 train_c3.csv 中。

将（3）-（4）小题的代码截图复制粘贴至客户端桌面【M2-T1-SUBT2-提交结果3.docx】中对应的任务序号下。

（二）任务二: 数据标注

对上述train_c3.csv数据进行标注，判断客户是否会购买银行的产品，具体的标注规则如下:

（1）如果“subscribe”列数据为1,则数据标注为‘yes’;

（2）如果“subscribe”列数据为0,则数据标注为‘no’;

标注好的数据存储为列‘subscribe’并和train.csv数据合并存入result.csv。

将代码截图复制粘贴至客户端桌面【M2-T2-SUBT1-提交结果1.docx】中对应的任务序号下。

（三）任务三: 数据清洗

1.子任务一: HDFS 文件上传下载

本任务需要使用 Hadoop，HDFS 命令，已安装 Hadoop 及需要配置前置环境，具体要求如下:

（1）在 Master 中的/root/目录下新建一个文件夹:result，将创建文件夹命令与结果截图粘贴至客户端桌面【M2-T3-SUBT1-提交结果 1.docx】中对应的任务序号下;

（2）使用 HDFS 命令，将 Master 下: /root 目录下新建的文件夹: result 上传到 HDFS 指定目录下: /根目录下; 并且使用 HDFS 命令查看目录; 将 HDFS 上传，查看命令截图粘贴至客户端桌面【M2-T3-SUBT1-提交结果 2.docx】中对应的任务序号下;

（3）使用 HDFS 命令，将 HDFS 目录下的/result 文件夹下载到 Master 指定目录下:根目录下; 将下载文件夹命令与结果截图粘贴至客户端桌面【M2-T3-SUBT1-提交结果 3.docx】中对应的任务序号下。

2.子任务二：处理异常数据

打开 ZZ052-7-M2-T3-SUBT2 文件夹，文件夹中包含 sku_info.csv 文件。sku_info.csv 文件存储了电商互联网平台上收集的商品数据，数据中有以下内容:

id: 主键非空，bigint 类型，长度为 20

spu_id: spuId，varchar 类型，长度 20

price: 价格，decimal 类型，长度 10

sku_name: 商品名称，varchar 类型，长度 200

sku_desc: 商品描述，varchar 类型，长度 2000

weight:重量，decimal 类型，长度 10

tm_id:品牌，bigint 类型，长度 20

11

category3_id:三级分类，bigint 类型，长度 20

sku_default_img:默认显示图片，varchar 类型，长度 200

create_time:创建时间，datetime 类型，长度 0，格式为 yyyy-MM-dd HH:mm:ss

编写 MapReduce 程序，实现以下功能：清除日志中字段长度比 11 小的日志记录，输出文件到 HDFS；在控制台按顺序打印输出前 20 条数据，将结果截图粘贴至客户端桌面【M2-T3-SUBT2-提交结果 1.docx】中对应的任务序号下。

3.子任务三：数据统计

打开 ZZ052-7-M2-T3-SUBT3 文件夹，文件夹中包含 user_info.csv 文件。user_info.csv 文件存储了电商互联网平台上收集的用户数据，数据中有以下内容:

id：主键非空，bigint 类型，长度为 20

login_name：用户名，varchar 类型，长度 200

nick_name：用户昵称，varchar 类型，长度 200

passwd：密码，varchar 类型，长度 200

name：姓名，varchar 类型，长度 200

phone_num:手机号，varchar 类型，长度 200

email:邮箱，varchar 类型，长度 200

head_img:头像，varchar 类型，长度 200

user_level:用户级别，varchar 类型，长度 200

birthday:用户生日，date 类型，长度 0，格式为

12

YYYY-MM-DD

gender:性别，varchar 类型，长度 1

create_time:创建时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

operate_time:修改时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

编写 MapReduce 程序，实现以下功能:对于 gender 这一字段统计电商消费人数男女数量，在控制台输出男女各多少人，将结果截图粘贴至客户端桌面【M2-T3-SUBT3-提交结果 1.docx】中对应的任务序号下。

四、模块三：业务分析与可视化

（一）任务一: 数据可视化

1.子任务一: 数据分析

打开 ZZ052-7-M3-T1-SUBT1 文件夹,文件夹中包含了 ANALYSE.xlsx 文件。对 ANALYSE.xlsx 文件中的数据，使用电子表格软件进行查询统计并使用图表进行展示。

(1)如参考截图所示,根据 ANALYSE.xlsx 文件的数据，使用电子表格软件统计岗位数量前十的城市，并以柱状图展示:

①统计每个“城市”的“岗位数量”;

②对“城市”的“岗位数量”进行降序排列;

13

③取“城市”的“岗位数量”前十使用柱状图进行显示。

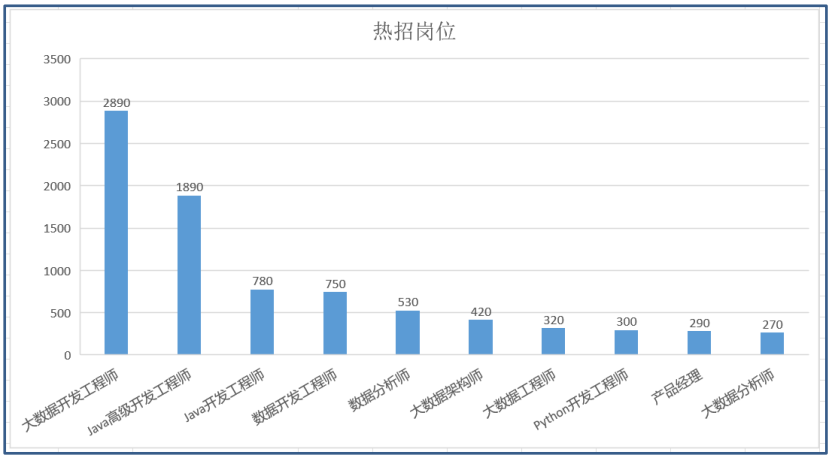

图 1 “岗位数量”结果参考示意图

(2)如参考截图所示,根据 ANALYSE.xlsx 文件的数据,使用电子表格软件统计各学历岗位数量占比,并以饼图展示:

①统计每种“学历”的“岗位数量”;

②根据每种“学历”的“岗位数量”计算每种学历岗位数的百分比;

③使用饼图展示每种学历岗位数的百分比。

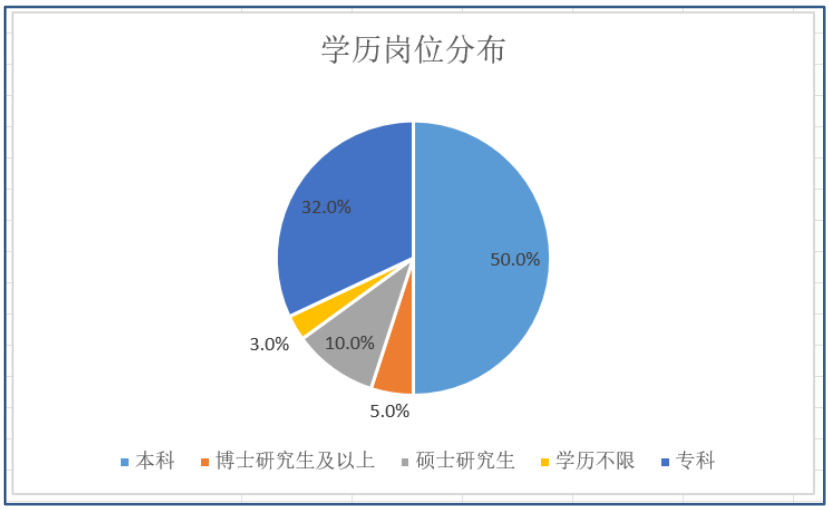

图 2 每种学历岗位数的百分比参考示意图

14

2.子任务二：数据可视化

打开 ZZ052-7-M3-T1-SUBT2 文件夹，文件夹中包含 jobSite 项目目录。打开 jobSite 项目，编写补充代码，实现 Web 网页形式对求职数据进行可视化展示。:

（1）如参考截图所示,根据 jobSite/data/data.js 文件中 hotskill 对象中的数据，补充完整 jobSite/js/chat.js 文件中 getHotskill()函数的代码，实现“热门技术”柱状图显示:

①编写补充 yAxis 对象，获取 hotskill 数据，设置 y 轴显示类型为“类目轴”、设置坐标轴文字显示为白色、设置 y 轴显示数据为“热门技术”;

②编写补充 series 对象，获取 hotskill 数据，设置图表显示类型为柱状图、设置柱状图文字显示效果、设置填充图表数据为 hotskill;

③运行网页，对“热门技术”柱状图进行截图。

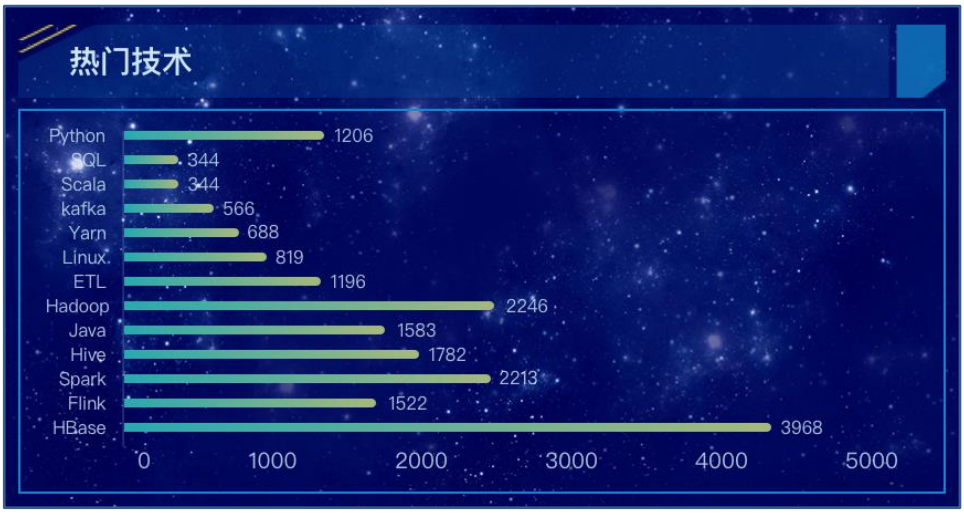

图 3 “热门技术”的柱状参考示意图

15

（2）如参考截图所示,根据 jobSite/data/data.js 文件中 salary 对象中的数据，补充完整 jobSite/js/chat.js 文件中 getSalaryData()函数的代码，实现“学历分布”饼图显示:

①编写补充 legend 对象，为饼图添加图例。设置图例的朝向为 垂直显示，设置图例在 X 轴方向上的位置为右、设置图例上显示的文字信息为“学历分布”、设置图例文字颜色为白色。

②编写补充 series 对象,设置图表的标题和图表类型，设置饼图半径为['20%', '55%']，设置饼图样式：饼图份例圆角度数为“4”将 salary 对象中的数据设置为饼图显示数据。

③运行网页，对“学历分布”饼图进行截图。

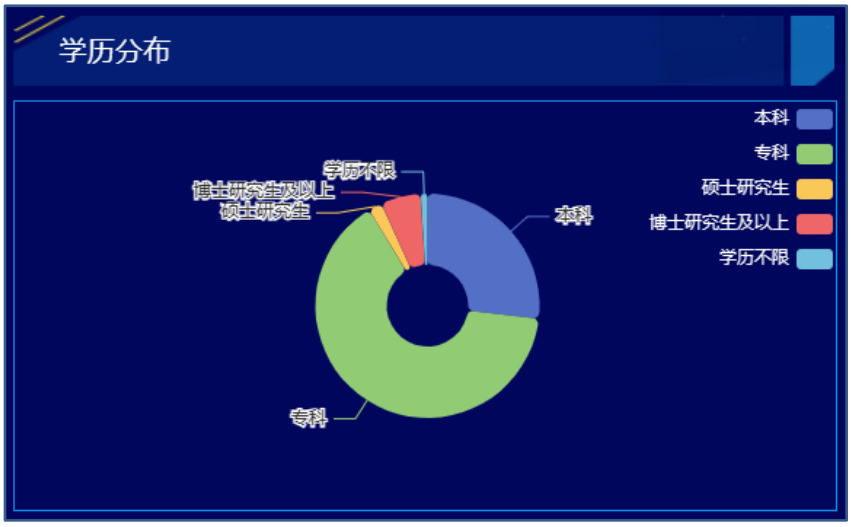

图 4 “学历分布”的饼状参考示意图

将该 2 个可视化图表分别截图复制粘贴至客户端桌面【M3-T1-SUBT2-提交结果 1.docx】中对应的任务序号下。

16

（二）任务二：业务分析

1.子任务一：业务分析

根据上述生成的train_c1.csv数据，分析不同职业的客户购买银行产品意向，并将分析结果用python绘制条形图。将图表截图复制粘贴至客户端桌面【M3-T2-SUBT1-提交结果1.docx】中对应的任务序号下，并在其下方编写发展趋势分析。

2.子任务二：报表分析

根据上述生成的 train_c1.csv 文件，通过 python 生成报表信息方便银行在后续服务中进行优化，及时准确的把握市场行情，具体要求如下：画出所有年龄（age）、随着上一次联系的时长（秒）（duration）、活动期间联系客户的次数（campaign）三个特征的直方图和概率密度图。将图表截一张图复制粘贴至客户端桌面【M3-T2-SUBT2-提交结果 1.docx】中对应的任务序号下。

17