2024年广西职业院校技能大赛
中职组《大数据应用与服务》赛项
竞赛样题

模块一：平台搭建与运维

利用竞赛平台进行大数据系统的安装和配置、数据库的安装和使用、平台运维、数据库运维等。

任务一：大数据平台搭建

使用SSH客户端通过SSH访问竞赛平台上的操作系统容器，基于竞赛平台进行伪分布式模式Hadoop的搭建和管理。相关安装文件在容器“/opt”目录下，请选择对应的安装包进行安装，用不到的可忽略。

1.在容器中执行命令，创建Hadoop安装目录“/data/hadoop”，执行ls命令查看创建的目录。

2.在容器中执行tar命令，将容器“/opt”目录下Hadoop安装文件解压到容器中“/data/hadoop”目录，执行ls命令查看解压后的文件。

3.配置Hadoop环境变量并使其生效,配置完毕后,在容器中执行“hadoop version”命令，查看Hadoop版本。

4.在容器中配置伪分布式模式Hadoop，并执行命令，格式化NameNode。

5.在容器中执行命令，启动HDFS。

6.在容器中执行jps命令，查看容器中的进程。

任务二：数据库配置维护

使用数据库客户端工具访问竞赛平台上的数据库容器，基于竞赛平台进行数据迁移和备份还原。

1.使用数据库工具，将MySQL中task1数据库的t_house_renting表数据迁移到数据库task1的house_renting表。

2.使用数据库工具，将MySQL中task1的house_renting表数据迁移到Excel文件house_renting.xls中。

3.使用SSH管理工具，执行SQL语句，备份表t_house_renting表到容器的“/opt/db/data/tab_bak”目录。

4.使用SSH管理工具,执行SQL语句,清空task1中表t_house_renting的数据。

5. 使用 SSH 管理工具，执行 SQL 语句，利用容器中“ /opt/db/data/tab_bak ”目录下的备份文件还原表 task1.t_house_renting。

模块二：数据获取与处理

本模块针对租房网站的数据进行采集、标注与处理。使用 Python 程序进行网站数据的读取与解析。使用 sql 对采集的租房数据进行清洗和标注。使用 Spark 编程读取提供的数据库表中的数据，按要求进行数据的预处理，并将处理完成的数据保存到数据库表中。

任务一中使用的网页和代码在素材文件夹中提供。

任务二中使用的租房数据文件在素材文件夹中提供。

任务三中需要处理的数据位于数据库 task2 中 house_lg 表，处理后的结果保存到数据库 task2 下 house_lg_op2 表中。建库脚本和表结构说明在素材文件夹中提供。

任务一：数据采集

任务要求：

打开 ZZ40-M2-T1 文件夹，文件夹中包含 parse_house.py 文件。house_renting.html 是通过爬虫爬下来的租房列表内容。parse_house.py 为 Python 脚本文件，程序读取 house_renting.html，使用 lxml 对网页进行解析，提取相应的租房列表数据，并将结果输出。

1.补全 parse_house.py 中【1】代码，配置公共资源地址 url。

2.使用浏览器打开 house_renting.html 网页文件，通过“审查”工具进行网页结构分析。

3.补全 parse_house.py 中【2】代码，实现获取租房 div 列表。

4.补全 parse_house.py 中【3】~【4】代码，实现“text”和“维护时间”文本内容提取。

5.运行 parse_house.py 脚本，完成租房列表的解析。

任务二：数据标注

打开 ZZ40-M2-T2 文件夹，请使用数据库工具导入 house_renting.xlsx 文件，按照下面的要求，进行数据处理。

1.利用数据库工具导入 house_renting.xlsx 文件。

2.使用 sql 语句将数据中“特点”为空的和“维护时间”超过三个月(包

含 3 个月前维护)的数据删除，并将数据导出到 house_renting_op.csv。

3.使用 sql 语句增加“方式”列，根据“名称”列的值来标注，如果名称中包含“整租”则标为“Z”，包含“合租”则标为“H”，并将数据导出到 house_renting_op2.csv。

任务三：Spark 数据处理

编辑赛项中提供的 ZZ4O-M2-T3/HOUSERENTINGOP 数据预处理程序，该程序使用 Spark 计算框架对租房数据进行预处理，请完成指定操作后在本地运行该程序。数据来自 MYSQL 数据库 task2 的 house_renting 表，预处理结果保存到 MYSQL 数据库 task2 下的 house_renting_op1 和 house_renting_op2 表中。

1.打开 ZZ4O-M2-T3/HOUSERENTINGOP/house_renting.py 文件，根据比赛分配的账号配置该文件下的数据库连接信息：server、port、user、password。

2.打开 ZZ4O-M2-T3/HOUSERENTINGOP/house_renting.py 文件,补充第 46 行代码，调用 SiteUdf 函数实现‘名称’数据归一化。

3.打开 ZZ4O-M2-T3/HOUSERENTINGOP/house_renting.py 文件,补充第 73 行代码,将预处理结果存入 MySQL 数据库 task2 的 house_renting_op2 表中。

4.运行程序。

5.使用数据库工具浏览数据库 task2 的 house_renting_op1 和 house_renting_op2 表，检查数据是否插入成功。

模块三：业务分析与可视化

对不同形式的求职数据进行分析和可视化，数据形式包括数据库表数据和 Web 程序数据。

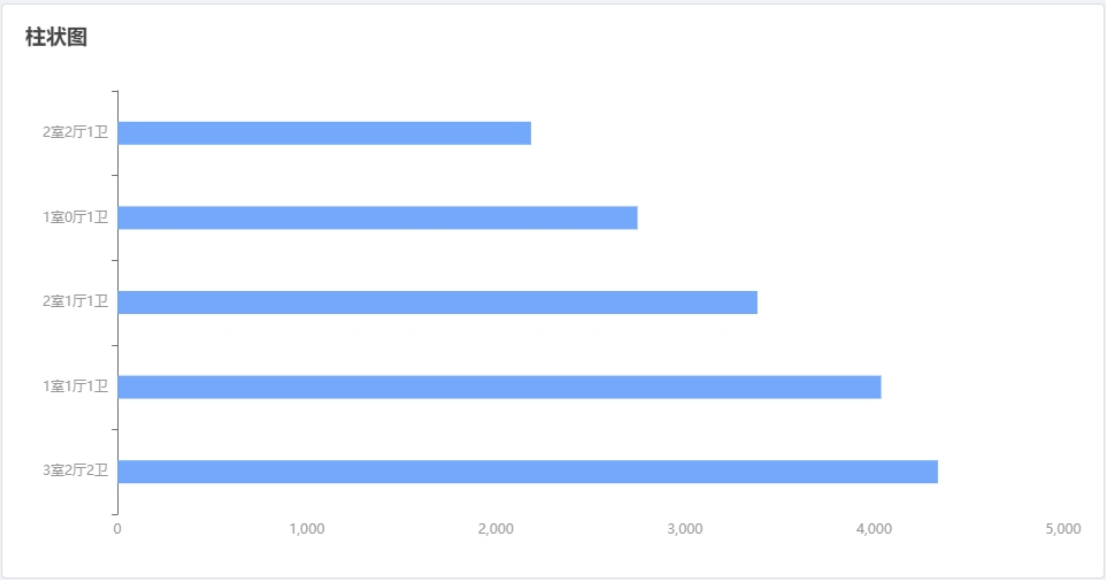
数据库表数据采用 MySQL 数据库进行存储，提供已建好的数据库表，使用数据库管理工具，运行 SQL 语句进行查询统计。Web 程序数据在 Web 程序代码中，使用 Web 前端编程技术补充 Web 程序代码，实现数据可视化网页。

任务一：Web 可视化

子任务 1：柱状图数据分析和可视化

使用大数据应用与服务平台的数据分析与可视化工具或者打开 ZZ4O-M3-T1 文件夹,文件夹中包含 visualization 项目目录。打开 visualization 项目,编写补充代码,实现 Web 网页形式对房型数量前五的房型可视化展示。

将柱状图截图，截图参考如下：



根据 visualization/data/data.js 文件中 barData 对象中的数据，补充完整 visualization/js/chat.js 文件中 getBarChart()函数的代码，实现“新房型数量前五的房型柱状图”显示：

1.编写补充 yAxis 对象，获取 barData 数据，设置 y 轴显示类型为“类目轴”、设置坐标轴文字颜色值为：#999999，大小为：12、设置坐标轴在 grid 区域中的分隔线颜色为：#CAD3E0，线的类型为：点虚线、设置 y 轴显示数据为“房型数量前五房型名称”。

2.编写补充 series 对象，获取 barData 数据，设置图表显示类型为柱状图、设置柱条的宽度为 20，背景颜色为：rgba(180，180，180，0.2)、设置填充图表数据为:房型数量。

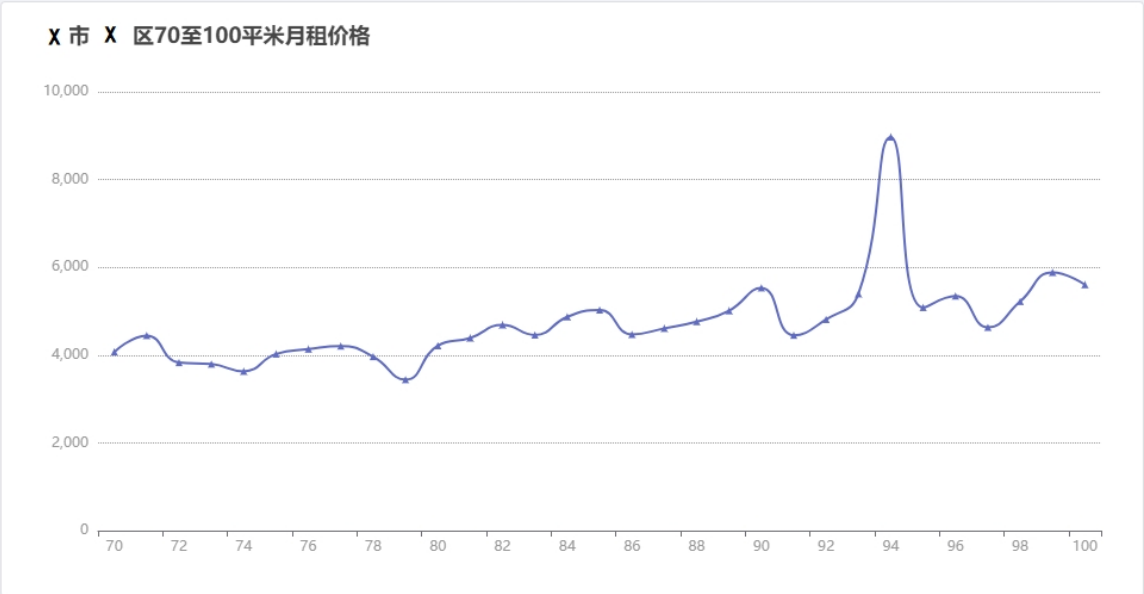
3.运行网页，附上“房型数量前五房型柱状图”截图与相关代码截图。

子任务 2：折线图数据分析和可视化

任务要求：

使用大数据应用与服务平台的数据分析与可视化工具或者打开 ZZ40-M3-T2 文件夹，文件夹中包含 visualization 项目目录。打开 visualization 项目，编写补充代码，实现 Web 网页形式对 X 市 X 区 70 至 100 平米月租价格可视化展示。

（1）将折线图截图，截图参考如下：



根据 visualization/data/data.js 文件中 lineData 对象中的数据，补充完整 visualization/js/chat.js 文件中 getLineChart()函数的代码，实现“X 市 X 区 70 至 100 平米月租价格折线图”显示：

1.编写补充 tooltip 对象，获取 lineData 数据，设置提示框组件的触发类型为坐标轴触发、设置指示器类型为：直线指示器、设置提示框浮层的文字颜色：##666666，字体大小为：12

2.编写补充 xAxis 对象，获取 lineData 数据，xAxis、设置坐标文字显示为:#99999，文字大小设置为：12、设置 X 轴显示坐标为“租房面积”。

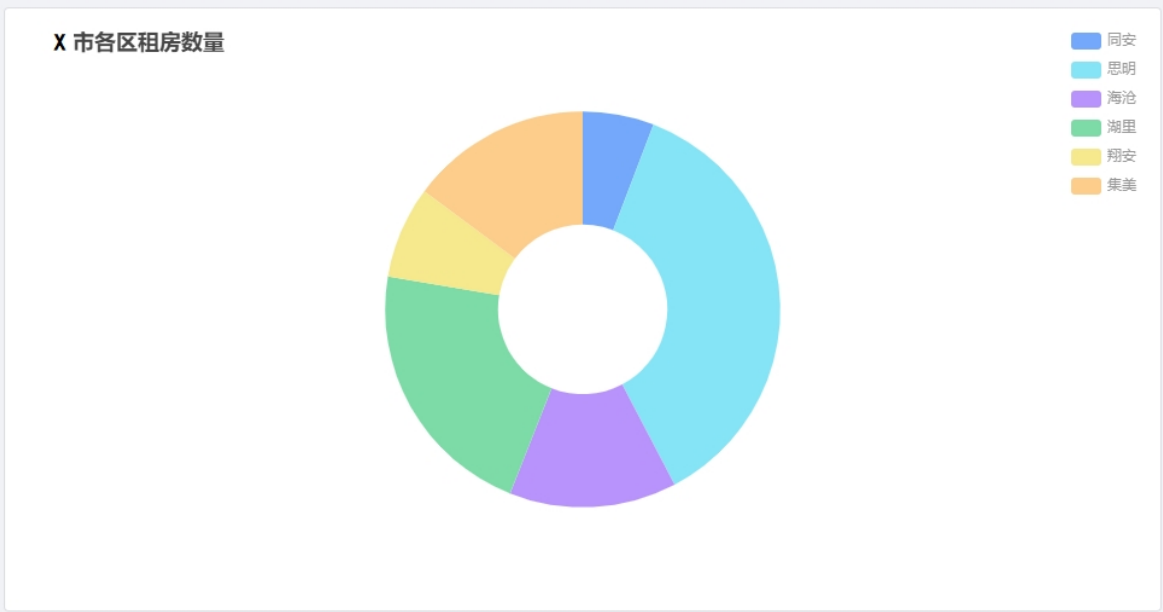
3.编写补充 series 对象，获取 lineData 数据，设置图表显示类型为‘line’、设置线条显示平滑,标记大小为 6 的三角形、设置折线图文字显示(将文字颜色设置为#999999、文字大小设置为：12)、将“X 市 X 区 70 至 100 平米月租价格”对象中的数据设置为折线显示数据。

运行网页，附上“X 市 X 区 70 至 100 平米月租价格折线图”截图与相关代码截图。

子任务 3：饼图数据分析和可视化

使用大数据应用与服务平台的数据分析与可视化工具或者打开 ZZ40-M3-T3 文件夹，文件夹中包含 visualization 项目目录。打开 visualization 项目，编写补充代码，实现 Web 网页形式对 X 市各区租房数量进行可视化展示。

将饼图截图，截图参考如下：



根据 visualization/data/data.js 文件中 pieData 对象中的数据，补充完整 visualization/js/chat.js 文件中 getPieChart()函数的代码，实现“X 市各区租房数量饼图”显示：

1.编写补充 legend 对象，获取 pieData 数据，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：六个区名称、设置图例文字颜色为：#999999，大小为：12。

2.编写补充 series 对象，获取 pieData 数据，xAxis、设置图表的标题和图表类型、设置饼图半径为['30%', '70%']、设置饼图高亮状态，标签文字颜色：#999999，大小：24，居中显示、将‘六个城市的在售房子套数’对象中的数据设置为饼图显示数据。

运行网页，附上“X 市各区租房数量”截图与相关代码截图。

任务二：业务分析

子任务 1：SQL 语句业务分析

打开数据库管理工具，使用 SQL 语句对 mysql 数据库下 task3 的 house_renting 表中的数据进行查询统计。house_renting 表结构参考 ZZ40-M3-T4 目录下“数据库表结构.docx”文档。

1.根据 house_renting 表的数据，使用 SQL 语句查询统计租房价格的具体数据，并生成视图：

1) 根据位置和价格计算每个位置的平均租房价格。

2) 取平均租房价格最高的三个进行显示。

3) 根据查询到的数据生成视图“AVGPRICE_LOCATION_3”。

2.根据 house_renting 表的数据，使用 SQL 语句查询统计租房信息：

1) 根据价格对所有房子进行升序排列，并筛选出价格最低的前五个。

2) 筛选出面积大于 100.00㎡的房子，并按照价格进行排序。

3.价格分析：

3) 根据房子价格排名，分析出影响房子价格的正向因素。

4) 通过对房子价格最低的五个房子各个维度的分析，写出影响价格的负面因素。

模块 4：职业素养（5 分）

中职大数据比赛职业素养评分关注团队合作、创新、问题解决、沟通、项目管理、职业操守等方面。参赛者需展现学科知识运用、自我学习、文档撰写能力。全面素养将在比赛中为团队成功贡献关键因素。

任务要求：

1.团队合作：能力在团队中合作协调，有效沟通，共同完成项目任务。评价团队成员之间的合作默契、协同工作的能力。

2.创新能力：考察参赛者在解决问题时的创造性思维和创新能力，包括提出独特的解决方案、采用新方法和技术等。

3.问题解决能力：能够迅速识别问题，采用合适的方法解决问题，具备分析和解决实际问题的能力。

4.沟通能力：能够清晰、准确地表达自己的观点，有效地与团队成员和评审沟通，包括书面报告、口头演讲等。

5.文档撰写能力：能够撰写清晰、完整、规范的文档。