

2024 年河北省职业院校技能大赛

大数据应用与服务（中职组）

赛项样题

项目背景

近年来随着社会经济的快速发展，百姓生活水平的不断提高，外出旅游成为很多人生活的热门选择，如何促进旅游业的发展成为各级政府高度重视的工作。为了更好地统筹管理城市的旅游资源，某省的旅游管理部门采集了本省若干城市的酒店经营数据和用户评论数据，其中酒店经营数据包括日期、城市、酒店名称、酒店星级、酒店当天预定房间数、酒店当天入住客户数、酒店当天最高房价和酒店当天最低房价等字段，这些数据保存到文件 `hotel.csv` 中。用户评论数据包括日期、城市、酒店名称、住客评分、评论内容等字段，这些数据保存到文件 `comments.csv` 中。

你作为技术人员，需要通过数据采集清洗、数据标注、数据分析、数据可视化、业务分析等步骤对酒店经营数据和用户评论数据进行处理，从而为政府制定旅游发展的政策提供决策依据。请按照下面的要求完成相关任务。

模块一：平台搭建与运维

任务一：大数据平台搭建

子任务 1 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)，将命令和结果复制粘贴至对应报告中；

(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，将命令和结果复制粘贴至对应报告中；

(3) 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，将命令和结果复制粘贴至对应报告中；

(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，将命令和结果复制粘贴至对应报告中；

(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，将命令和结果复制粘贴至对应报告中。

子任务 2 Hive 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-hive-3.1.2-bin.tar.gz、mysql-connector-java-5.1.37.jar 安装包解压到/opt/module 目录下，将命令和结果复制粘贴至对应报告中。

2)设置 Hive 环境变量，并使环境变量生效，执行命令 hive --version 将命令和结果复制粘贴至对应报告中。

3)完成相关配置并添加所依赖包，将 MySQL 数据库作为 Hive 元数据库。初始化 Hive 元数据，并通过 schematool 相关命令执行初始化，将命令和结果复制粘贴至对应报告中。

任务二：数据库配置维护

子任务 1 在数据库中创建表

本任务在 MySQL 中创建表 t_comment 和表 t_hotel，并将用户评论数据 comments.csv 和酒店经营数据 hotel.csv 分别导入到表 t_comment 和表 t_hotel 中。具体要求如下：

1、创建用户评论表 t_comment，表 t_comment 的字段定义如下：

| 字段 | 类型 | 说明 | 备注 |
|--------------|---------|------|----|
| comment_date | date | 日期 | |
| city | varchar | 城市 | |
| hotel_name | varchar | 酒店名称 | |
| score | double | 住客评分 | |
| content | varchar | 评论内容 | |

2、在 MySQL 中将 comments.csv 的数据导入表 t_comment。

3、创建酒店经营数据表 t_hotel,表 t_hotel 的字段定义如下：

| 字段 | 类型 | 说明 | 备注 |
|--------------|---------|----|----|
| current_date | date | 日期 | |
| city | varchar | 城市 | |

| | | | |
|---------------------|---------|-----------|--|
| hotel_name | varchar | 酒店名称 | |
| hotel_star | varchar | 酒店星级 | |
| rooms_booked | int | 酒店当天预定房间数 | |
| customers_checkedin | int | 酒店当天入住客户数 | |
| highest_price | int | 酒店当天最高房价 | |
| lowest_price | int | 酒店当天最低房价 | |

4、在 MySQL 中将 hotel.csv 的数据导入表 t_hotel。

5、将以上 SQL 语句和运行结果复制粘贴至对应报告中。

子任务 2 使用 SQL 查询数据

本任务具体要求如下：

- 1、查询指定酒店的评论数量。
- 2、查询指定酒店的住客评分的平均值。
- 3、查询每个城市的酒店数量。
- 4、查询指定酒店的最高房价和最低房价。
- 5、将以上 SQL 语句和运行结果复制粘贴至对应报告中。

模块二：数据获取与处理

任务一：数据获取与清洗

子任务 1 对空字段数据进行处理

1、使用 python 读取 comments.csv 文件，将字段“酒店名称”为空的数据删除，并打印输出删除条目数，将打印内容粘贴至对应报告中，打印内容格式如下：

```
=== “删除酒店名称为空的数据共***条” ===
```

2、将字段“酒店名称”非空的数据保存到 comments1.csv 文件。

3、将符合题目要求的代码答案和 comments1.csv 的前 10 条记录数据复制粘贴至对应报告中。

子任务 2 对异常字段数据进行处理

住客评分的取值范围为[0,5]，其中 5 表示评价最高，0 表示评价最低。如果住客评分超出此取值范围的，都视为异常数据。本任务使用 python 读取 hotel.csv 文件的数据，将字段“住客评分”异常的数据删除，并打印输出删除条目数，将打印内容粘贴至对应报告中，打印内容格式如下：

```
=== “删除住客评分异常的数据共***条” ===
```

任务二：数据标注

本任务根据酒店的评论数据对酒店的类型打上标签，并将标签数据保存到指定位置。系统提前设定用户评价活跃阈值，如酒店的用户评价数量大于用户评价活跃阈值，则将该酒店的类型标注为“热门”，否则将该酒店的类型标注为“普通”，具体要求如下：

1、编写 python 程序读取读取 comments.csv 的数据，统计每个酒店的用户评价数量。

2、比较酒店的评价数量和用户评价活跃阈值，给该酒店的类型打上指定的标签（热门/普通），然后将打上标签的数据保存到 comments_tag.csv 中，comments_tag.csv 的字段定义如下：

| 酒店名称 | 评论数量 | 酒店类型 |
|------|------|-------|
| | | 热门/普通 |
| | | |

任务三：数据统计

本任务使用 MapReduce 程序对酒店经营数据进行统计。

子任务 1 统计每个酒店的预订房间总数和入住客户总数

- 1) 将 hotel.csv 文件上传至 HDFS 目录/hotel 中。
- 2) 编译打包 MapReduce 程序，并将代码部署在 Hadoop 平台上运行，将程序运行结果保存到 HDFS 目录/result1 下。
- 3) 读取 HDFS 目录/result1 的数据，将该数据复制粘贴至对应报告中。

子任务 2 统计每个城市不同星级酒店的数量

- 1) 将 hotel.csv 文件上传至 HDFS 目录/hotel 中。
- 2) 编译打包 MapReduce 程序，并将代码部署在 Hadoop 平台上运行，将程序运行结果保存到 HDFS 目录/result2 下。
- 3) 读取 HDFS 目录/result2 的数据，将该数据复制粘贴至对应报告中。

模块三：业务分析与可视化

任务一：数据可视化

子任务 1 使用堆叠图展示城市星级酒店的数量

本任务使用堆叠图展示每个城市星级酒店的数量，本任务具体要求如下：

- 1) 读取 hotel.csv，使用 pandas 分别统计每个城市的三星级酒店、四星级酒店和五星级酒店的数量。
- 2) 使用 matplotlib 绘制堆叠图，堆叠图的标题为“各城市星级酒店的数量”，堆叠图的横坐标为城市名称，纵坐标为星级酒店数量。将可视化结果复制粘贴至对应报告中。

子任务 2 使用散点图展示各城市酒店入住客户总人数

将每个城市的所有酒店的入住客户的数量进行累加，就获得了每个城市入住客户的总人数。使用散点图展示不同城市入住客户的总人数，可以直观地对比这些城市的旅游接待能力，本任务具体要求如下：

1) 读取 `hotel.csv`，使用 `pandas` 统计每个城市的所有酒店的入住客户总人数。

2) 使用 `matplotlib` 绘制散点图，散点图的标题为“各城市酒店入住客户总人数”，将可视化结果复制粘贴至对应报告中。

子任务 3 使用柱状图展示酒店的评分数据

本任务使用柱状图展示酒店的评分数据，具体要求如下：

1) 读取 `hotel.csv`，使用 `pandas` 统计分别统计三星级酒店、四星级酒店和五星级酒店的住客评分的平均值。

2) 使用 `matplotlib` 绘制柱状图，柱状图的标题为“不同星级酒店的住客评分数据”，柱状图的横坐标分别为三星级酒店、四星级酒店和五星级酒店，纵坐标为星级酒店对应的住客评分的平均值。柱状图为横向布局，将可视化结果复制粘贴至对应报告中。

任务二：业务分析

子任务 1 分析影响酒店入住客户数量的因素有哪些

结合模块三的任务一制作的可视化效果图，说明影响酒店入住客户数量的因素有哪些，并就如何提高酒店入住率给出相应的措施和建议。

子任务 2 分析影响酒店评分的因素有哪些

结合本模块三的任务一制作的可视化效果图，说明影响酒店评分

的因素有哪些,并就如何提高酒店用户满意度和服务水平给出相应的措施和建议。