2023-2024 年广东省职业院校技能大赛

中职组大数据应用与服务赛项

样

题

4

一、背景描述

大数据时代背景下，人们生活习惯发生了很多改变。在传统运营模式中，缺乏数据积累，人们在做出一些决策行为过程中，更多是凭借个人经验和直觉，发展路径比较自我封闭。而大数据时代，为人们提供一种全新的思路，通过大量的数据分析得出的结果将更加现实和准确。平台可以根据用户的浏览，点击，评论等行为信息数据进行收集和整理。通过大量用户的行为可以对某一个产品进行比较准确客观的评分和评价，或者进行相应的用户画像，将产品推荐给喜欢该产品的用户进行相应的消费。

因数据驱动的大数据时代已经到来，没有大数据，我们无法为用户提供大部分服务，为完成互联网酒店的大数据分析工作，你所在的小组将应用大数据技术，通过 Python 语言以数据采集为基础，将采集的数据进行相应处理，并且进行数据标注、数据分析与可视化、通过大数据业务分析方法实现相应数据分析。运行维护数据库系统保障存储数据的安全性。通过运用相关大数据工具软件解决具体业务问题。你们作为该小组的技术人员，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：基础环境准备

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）配置三个节点的主机名，分别为 master、slave1、slave2，然后修改三个节点的 hosts 文件，使得三个节点之间可以通过主机名访问，在 master 上将执行命令 cat /etc/hosts 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（2）将 /opt/software 目录下将文件 jdk-8u191-linux-x64.tar.gz 安装包（若 slave1、slave2 节点不存在以上文件则需从 master 节点复制）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【提交结果.docx】中对应的任务序号下；

（3）在 /etc/profile 文件中配置 JDK 环境变量 JAVA_HOME 和 PATH 的值，并让配置文件立即生效，将在 master 上 /etc/profile 中新增的内容复制并粘贴至【提交结果.docx】中对应的任务序号下；

（4）查看 JDK 版本，检测 JDK 是否安装成功，在 master 上将执行命令 java -vserion 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（5）创建 hadoop 用户并设置密码，为 hadoop 用户添加管理员权限。在 master 上将执行命令 grep 'hadoop' /etc/sudoers 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（6）关闭防火墙，设置开机不自动启动防火墙，在 master 上将执行命令 systemctl status fireawlld 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（7）配置三个节点的 SSH 免密登录，在 master 上通过 SSH 连接 slave1 和 slave2 来验证。

2. 子任务二：Hadoop 完全分布式安装配置

本任务需要使用 root 用户和 hadoop 用户完成相关配置，使用三个节点完成 Hadoop 完全分布式安装配置。命令中要求使用绝对路径，具体要求如下：

（1）在 master 节点中的 /opt/software 目录下将文件 hadoop-3.3.6.tar.gz 安装包解压到 /opt/module 路径中，将 hadoop 安装包解压命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（2）在 master 节点中将解压的 Hadoop 安装目录重命名为 hadoop ，并修改该目录下的所有文件的所属者为 hadoop，所属组为 hadoop，将修改所属者的完整命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（3）在 master 节点中使用 hadoop 用户依次配置 hadoop-env.sh、core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml、masters 和 workers 配置文件，Hadoop 集群部署规划如下表，将 yarn-site.xml 文件内容复制并粘贴至【提交结果.docx】中对应的任务序号下；

服务器	master	slave1	slave2
DHFS	NameNode		
HDFS	SecondaryNameNode		
HDFS	DataNode	DataNode	DataNode
YARN	ResourceManager		
YARN	NodeManager	NodeManager	NodeManager
历史日志服务器	JobHistoryServer		

（4）在 master 节点中使用 scp 命令将配置完的 hadoop 安装目录直接拷贝至 slave1 和 slave2 节点，将完整的 scp 命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（5）在 slave1 和 slave2 节点中将 hadoop 安装目录的所有文件的所属者为 hadoop，所属组为 hadoop。

（6）在三个节点的 /etc/profile 文件中配置 Hadoop 环境变量 HADOOP_HOME 和 PATH 的值，并让配置文件立即生效，将 master 节点中 /etc/profile 文件新增的内容复制并粘贴至【提交结果.docx】中对应的任务序号下；

（7） 在 master 节点中初始化 Hadoop 环境 namenode，将初始化命令及初始化结果（截取初始化结果日志最后 20 行即可）粘贴至【提交结果.docx】中对应的任务序号下；

（8） 在 master 节点中依次启动 HDFS、YARN 集群和历史服务。在 master 上将执行命令 jps 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下；

（9） 在 slave1 查看 Java 进程情况。在 slave1 上将执行命令 jps 的结果复制并粘贴至【提交结果.docx】中对应的任务序号下。

（二）任务二：数据库服务器的安装与运维

1. 子任务一：MySQL 安装配置

本任务需要使用 rpm 工具安装 MySQL 并初始化，具体要求如下：

（1） 在 master 节点中的 /opt/software 目录下将 MySQL 5.7.44 安装包解压到 /opt/module 目录下；

（2）在 master 节点中使用 rpm -ivh 依次安装 mysql-community-common、mysql-community-libs、mysql-community-libs-compat 、 mysql-community-client 和 mysql-community-server 包，将所有命令复制粘贴至【提交结果.docx】 中对应的任务序号下；

（3）在 master 节点中启动数据库系统并初始化 MySQL 数据库系统，将完整命令复制粘贴至【提交结果.docx】 中对应的任务序号下；

2.子任务二：MySQL 运维

本任务需要在成功安装 MySQL 的前提，对 MySQL 进行运维操作，具体要求如下：

（1）配置服务端 MySQL 数据库的远程连接，将新增的配置内容复制粘贴至【提交结果.docx】 中对应的任务序号下；

（2）配置 root 用户允许任意 IP 连接，将完整命令复制粘贴至【提交结果.docx】中对应的任务序号下；

（3）通过 root 用户登录 MySQL 数据库系统，查看 mysql 库下的所有表，将完整命令及执行命令后的结果复制粘贴至【提交结果.docx】中对应的任务序号下；

（4）输入命令以创建新的用户 hadoop，将完整命令及执行命令后的结果复制粘贴至【提交结果.docx】中对应的任务序号下；

（5）授予新用户查询数据和插入数据的权限，将完整命令及执行命令后的结果复制粘贴至【提交结果.docx】中对应的任务序号下；

（6）刷新权限，将完整命令及执行命令后的结果复制粘贴至【提交结果.docx】中对应的任务序号下。

3.子任务三：数据表的创建及维护

（1）根据以下数据字段在 bigdata 数据库中创建酒店表 （hotel）。酒店表字段如下：

字段	类型	中文含义	备注
id	int	酒店编号	
hotel_name	varchar	酒店名称	
city	varchar	城市	
province	varchar	省份	
level	varchar	星级	
room_num	int	房间数	
score	double	评分	
commnet_num	varchar	评论数	

（2）根据以下数据字段在 bigdata 数据库中创建评论表 （comment）。评论表字段如下：

字段	类型	中文含义	备注
id	int	评论编号	
name	varchar	酒店名称	
commentator	varchar	评论人	
score	double	评分	
comment_time	datetime	评论时间	
content	varchar	评论内容	

将这两个 SQL 建表语句分别复制粘贴至【提交结果.docx】中对应的任务序号下。

（3）根据已给到的 sql 文件将这两份数据导入任意自己创建的数据库中，并对其中的数据进行如下操作：

- 在 comment 表中将 id 为 30 的评分改为 5；
- 在 hotel 表中统计各城市的酒店总数；

  将这两个 SQL 语句分别复制粘贴至 【提交结果.docx】中对应的任务序号下。

4. 子任务四：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 ）在 master 节点将 /opt/software 目录下的 apache-hive-3.1.3-bin.tar.gz 安装包解压到 /opt/module 路径下， 将完整命令截图粘贴至【提交结果.docx】中对应的任务序号下；

（2）把解压后的 apache-hive-3.1.3-bin 文件夹更名为 hive-3.1.3，并修改该目录下的所有文件的所属者为 hadoop，所属组为 hadoop，将修改所属者的完整命令复制并粘贴至【提交结果.docx】中对应的任务序号下；

（3）配置 Hive 环境变量，并使环境变量生效，将新增的环境变量内容截图粘贴至【提交结果.docx】中对应的任务序号下；

（4）修改 hive-site.xml 配置文件，将 MySQL 数据库作为 Hive 元数据库，将配置 Hive 元数据库的相关内容截图粘贴至【提交结果.docx】中对应的任务序号下；

（6）将 /opt/software 目录下的 MySQL 数据库 JDBC 驱动 mysql-connector-j-8.2.0.jar 拷贝到必要的文件夹下，将完整命令截图粘贴至【提交结果.docx】中对应的任务序号下；

（7）初始化 Hive 元数据库，将初始化命令及结果截图粘贴至【提交结果.docx】中对应的任务序号下；

（8）启动 Hive，将命令输出结果截图粘贴至【提交结果.docx】中对应的任务序号下；

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

有一份酒店详情列表数据：酒店名称，地区，地址，卫生评分，服务评分，设施评分，位置评分，评价数，装修时间，房间类型，房价，经度，纬度，公司，出行住宿，校园生活。

并且已存入到 hotel.csv 文件中，请使用 pandas 读取 hotel.csv 并将数据集的前 10 行打印在 IDE 终端的截图复制粘贴至【提交结果.docx】 中对应的任务序号下。

2. 子任务二：使用 Python 进行数据清洗

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数据脱敏。

相关数据文件中已经包含了数据采集阶段从企业消费平台网站上爬取的数据集，其中包含了来自北京的多家酒店的信息，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酒店数据的清洗和整理。

请使用 pandas 库加载并分析相关数据集，根据题目规定要求使用 pandas 库实现数据处理，具体要求如下：

（1）删除地址为空的记录，并将结果存储为 cleaned_data_c1_N.csv，N 为删除的数据条数；

（2）删除评分中任一项（卫生评分、服务评分、设施评分、位置评分）小于 3 分的记录，并将结果存储为 cleaned_data_c2_N.csv，N 为删除的数据条数；

（3）将装修时间为空的记录设置为数据集中最常见的装修时间，并存储为 cleaned_data_c3_N.csv，N 为修改的数据条数；

（4）将房价超过数据集房价均值三倍的记录视为异常值删除，并将结果存储为 cleaned_data_c4_N.csv，N 为删除的数据条数；

（5）识别并删除重复记录，将结果存储为 cleaned_data_c5_N.csv，N 为删除的数据条数；

将该 5 个文件名截一张图复制粘贴至 【提交结果.docx】中对应的任务序号下。

（二）任务二：数据标注

1. 子任务一：房价分类标注

使用 Python 编写脚本，根据房价将酒店分为三类：“经济型”、“中档型”和“豪华型”。具体的分类要求如下：

（1）经济型：房价低于 500 元/晚；

（2）中档型：房价在 500 元至 1500 元/晚之间；

（3）豪华型：房价超过 1500 元/晚；

在数据集中新增一列“房价分类”，根据上述标准对每家酒店进行分类标注，存入 hotel_mark_c1.csv 文件中。具体格式如下：

编号	酒店名称	地区	房价分类
1	北京瑜舍	朝阳区	豪华型

将 hotel_mark_c1.csv 打开后直接截图（不用下拉）复制粘贴至【提交结果.docx】中对应的任务序号下。

2. 子任务二：综合评级标注

使用 Python 编写脚本，基于卫生评分和服务评分，为每家酒店分配一个综合评级。具体的分类要求如下：

（1）五星级：卫生评分和服务评分的平均值≥4.5；

（2）四星级：卫生评分和服务评分的平均值在 4.0 至 4.49 之间；

（3）三星级：卫生评分和服务评分的平均值在 3.5 至 3.99 之间；

（4）二星级：卫生评分和服务评分的平均值在 3.0 至 3.49 之间；

（5）一星级：卫生评分和服务评分的平均值＜3.0；

在数据集中新增一列“综合评级”，根据上述标准为每家酒店进行评级标注。存入 hotel_mark_c2.csv 文件中。具体格式如下：

编号	酒店名称	装修时间	综合评级
1	北京瑜舍	2008	五星级

将 hotel_mark_c2.csv 打开后直接截图（不用下拉）复制粘贴至【提交结果.docx】中对应的任务序号下。

（三）任务三：数据统计

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 HDFS 目录下新建目录 /file2_1，将新建目录的完整命令粘贴至【提交结果.docx】中对应的任务序号下；

（2）修改权限，赋予目录 /file2_1 最高 777 权限，将修改目录权限的完整命令粘贴至【提交结果.docx】对应的任务序号下；

（3）下载 HDFS 新建目录 /file2_1，到本地容器 master 指定目录 /tmp 下，将完整命令粘贴至【提交结果.docx】中对应的任务序号下。

2. 子任务二：计算输入文件中的单词数

本任务需要使用 Hadoop 默认提供的 wordcount 示例来完成单词数统计任务，具体要求如下：

（1）在 HDFS 上创建 /user/hadoop/input 目录；

（2）在 master 节点将 /var/log/dmesg 文件上传到 HDFS 的 /user/hadoop/input 目录下；

（3）使用 Hadoop 中提供的 wordcount 示例对 HDFS 上的 dmesg 文件进行单词统计，并将统计结果存储到 HDFS 的 /user/hadoop/output 目录下；

（4）查看 HDFS 中的 /user/hadoop/output 单词数统计结果并将结果前十行截图粘贴至【提交结果.docx】中对应的任务序号下。

3. 子任务三：数独解算器

本任务需要使用 Hadoop 默认提供的 sudoku 示例来完成数独题目的解题任务，具体要求如下：

（1）使用 Hadoop 提供的 sudoku 示例计算以下数独题目：

```
8 5 ? 3 9 ? ? ? ?
? ? 2 ? ? ? ? ? ?
? ? 6 ? 1 ? ? ? 2
? ? 4 ? ? 3 ? 5 9
? ? 8 9 ? 1 4 ? ?
3 2 ? 4 ? ? 8 ? ?
9 ? ? ? 8 ? 5 ? ?
? ? ? ? ? ? 2 ? ?
? ? ? ? 4 5 ? 7 8
```

（2）将数独解题结果截图粘贴至【提交结果.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：数据分析与可视化

1. 子任务一：数据分析

数据分析是理解和解释数据的过程，对于酒店行业来说，它可以揭示客户偏好、市场趋势和运营效率。在这个任务中，我们将运用 Python 对酒店数据 hotel.csv 进行深入分析，以揭示行业的关键趋势和洞察。参赛者需要运用 Python 的数据处理和分析库，如 Pandas 来完成以下任务：

（1）分析各地区酒店的分布情况，统计每个地区的酒店数量，进行倒序排序展示前两名；

（2）计算不同装修时间的酒店的平均设施评分；

（3）计算每个地区的平均房价，进行倒序排序展示前三名；

（4）统计数据集中每种房间类型的出现频次，进行正序排序展示前两名；

（5）筛选出卫生评分和服务评分均高于 4.5 的酒店，并统计这些酒店的数量；

将该 5 个统计结果在 IDE 的控制台中打印并分别截图复制粘贴至【提交结果.docx】中对应的任务序号下。

2. 子任务二：数据可视化

在这个任务中，参赛者将使用 pyecharts 库来创建直观、互动的图表。这些图表将帮助揭示数据中的关键模式和趋势。 具体要求如下：

（1）使用柱状图展示各个地区的酒店数量，柱状图中的每个柱子代表一个地区，高度代表该地区的酒店数量；

（2）创建条形图比较不同地区酒店的平均卫生评分和服务评分，条形图中横轴表示评分，纵轴表示地区，每个地区有两个条形，分别表示卫生和服务评分；

（3）使用雷达图展示任意一间酒店在卫生、服务、设施和位置评分的综合表现，雷达图中的每个轴代表一个评分指标（卫生、服务、设施、位置），每个酒店表现为雷达图上的一个闭合路径；

（4）制作散点图探索房价与评价数量之间的关系，散点图中的横轴为房价，纵轴为评价数量，每个点代表一家酒店；

将该 4 个可视化图表分别截图复制粘贴至【提交结果.docx】中对应的任务序号下。

（二）任务二：业务分析

在酒店行业中，准确理解客户需求、市场趋势和经营效率是至关重要的。通过分析酒店的运营数据，可以获得提升服务质量和优化管理策略的宝贵信息。本任务中使用 Python 对酒店数据进行简单的业务分析，目的是识别酒店的主要客户群体特征，并提出基于数据的简单营销策略。

使用提供的酒店数据集，计算以下指标：

（1）平均房价：计算数据集中所有酒店的平均房价。

（2）最受欢迎的房间类型：根据评价数量，确定哪种房间类型最受欢迎。

（3）高评分酒店特征：识别评分在前 25%的酒店共有的特征（如房价范围、位置等）。

根据上述分析结果，撰写一段简短的描述，提出两条针对酒店的营销策略建议。将内容复制粘贴至 【提交结果.docx】中对应的任务序号下。