

广东省职业院校技能大赛

（中职组）大数据应用与服务赛项

2024-2025学年 大数据应用与服务赛项样题五

赛项编号：ZZ052

整理时间：2026年8月

场景：互联网酒店数据分析

大数据时代背景下，平台根据用户的浏览、点击、评论等行为信息数据进行收集整理，通过大量用户行为对产品进行评分评价和用户画像。为完成互联网酒店的大数据分析工作，需要应用大数据技术，通过Python语言以数据采集为基础，进行数据标注、数据分析与可视化、大数据业务分析，运行维护数据库系统保障数据安全。

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（root用户，命令要求绝对路径）

- （1）配置三个节点主机名 master、slave1、slave2，修改三节点 hosts 文件使节点间可通过主机名访问，在master上执行 cat /etc/hosts 并将结果粘贴至【提交结果.docx】对应任务序号下；
- （2）将 /opt/software 下 jdk-8u191-linux-x64.tar.gz 解压到 /opt/module（若slave1、slave2不存在则先从master复制），提交JDK解压命令；
- （3）在 /etc/profile 中配置 JAVA_HOME 和 PATH 并立即生效，提交master上新增内容；
- （4）执行 java -version 检测JDK安装成功，提交结果；
- （5）创建 hadoop 用户并设置密码、添加管理员权限，提交 grep 'hadoop' /etc/sudoers 结果；
- （6）关闭防火墙并设置开机不自动启动，提交 systemctl status firewalld 结果；
- （7）配置三个节点的 SSH 免密登录，在master上通过SSH连接slave1和slave2验证。

2. 子任务二：Hadoop 完全分布式安装配置

- （1）在master的 /opt/software 下将 hadoop-3.3.6.tar.gz 解压到 /opt/module，提交解压命令；
- （2）将解压目录重命名为 hadoop，并修改目录下所有文件所有者为 hadoop:hadoop，提交完整命令；
- （3）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、workers，提交 yarn-site.xml 内容；
- （4）使用scp命令将配置好的hadoop目录拷贝至slave1和slave2，提交完整scp命令；
- （5）slave1、slave2 修改hadoop目录所有文件所有者为 hadoop:hadoop；
- （6）三节点 /etc/profile 配置 HADOOP_HOME 和 PATH 并生效，提交master新增内容；
- （7）master初始化namenode，提交初始化命令及初始化结果日志最后20行；
- （8）依次启动HDFS、YARN集群和历史服务，提交master上 jps 结果；
- （9）在slave1执行 jps 查看Java进程，提交结果。

Hadoop集群部署规划：

服务器	master	slave1	slave2
HDFS-NameNode	√		
HDFS-SecondaryNameNode	√		
HDFS-DataNode	√	√	√
YARN-ResourceManager	√		
YARN-NodeManager	√	√	√

服务器	master	slave1	slave2
历史日志-JobHistoryServer	√		

（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

- （1）在master的 /opt/software 下将 MySQL 5.7.44 安装包解压到 /opt/module 目录；
- （2）使用 rpm -ivh 依次安装 mysql-community-common、mysql-community-libs、mysql-community-lib
s-compat、mysql-community-client、mysql-community-server，将所有命令粘贴至【提交结果.docx】；
- （3）启动并初始化MySQL数据库系统，提交完整命令。

2. 子任务二：MySQL 运维

- （1）修改MySQL配置文件启用远程连接，提交修改后的配置内容；
- （2）启用MySQL慢查询日志功能，设置慢查询阈值为2秒，提交完整的配置命令和验证命令结果；
- （3）创建名为 hadoop 的数据库用户，密码设置为Hadoop@2025，提交创建用户的完整命令及执行结果；
- （4）为 hadoop 用户授予 bigdata 数据库的SELECT、INSERT和UPDATE权限，提交授权命令及执行结果；
- （5）创建名为 monitor 的监控专用用户，仅授予PROCESS和SHOW
DATABASES权限，提交完整命令及执行结果；
- （6）使用mysqldump工具备份bigdata数据库，备份文件保存为bigdata_backup.sql，提交备份命令及执行
过程；
- （7）配置数据库字符集为utf8mb4，默认排序规则设置为utf8mb4_unicode_ci，提交配置命令和验证结果
；
- （8）创建定时备份脚本，实现每天凌晨2点自动备份bigdata数据库到/opt/backup目录，并只保留最近7天
的备份文件，提交完整脚本内容。

3. 子任务三：数据表的创建及维护

在bigdata数据库中创建酒店表和评论表：

- （1）hotel表字段：id(int, 酒店编号)、hotel_name(varchar, 酒店名称)、city(varchar, 城市)、provinc
e(varchar, 省份)、level(varchar, 星级)、room_num(int, 房间数)、score(double, 评分)、commnet_num(v
archar, 评论数)；
- （2）comment表字段：id(int, 评论编号)、name(varchar, 酒店名称)、commentator(varchar, 评论人)、sc
ore(double, 评分)、comment_time(datetime, 评论时间)、content(varchar, 评论内容)；
- （3）将两个SQL建表语句分别粘贴至【提交结果.docx】；
- （4）导入数据后操作：统计各省份的平均酒店评分，并将评分低于该省份平均分的酒店评分上调0.5分（
评分上限为5分）；查找出每个城市评论数最多的酒店及其详细信息；找出所有发表过3条及以上差评（评
分小于等于2分）的评论人，列出这些评论人的评论时间、评论内容及对应的酒店名称，按评论时间降序排
序。将这3个SQL语句分别粘贴。

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），已存入hotel.csv。使用pandas读取并将数据集前10行打印在IDE终端截图粘贴。

2. 子任务二：使用Python进行数据清洗

数据已脱敏处理。请使用 pandas 库加载并分析数据集，按要求完成清洗，各清洗结果文件以 cleaned_data_cN_N.csv 命名（N为删除/修改的数据条数）。

- （1）删除酒店名称为空的记录，存为 cleaned_data_c1_N.csv；
- （2）将评分数据中的空值替换为0分（卫生、服务、设施、位置评分），存为 cleaned_data_c2_N.csv；
- （3）删除评价数为0但任一评分项不为0的矛盾记录，存为 cleaned_data_c3_N.csv；
- （4）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，存为 cleaned_data_c4_N.csv；
- （5）删除地址字段完全为空的记录，存为 cleaned_data_c5_N.csv；
- （6）将房价为0或空值的记录删除，存为 cleaned_data_c6_N.csv；
- （7）统一房间类型的表述，将“标准间”“标准房”统一改为“标准双人房”，存为 cleaned_data_c7_N.csv；
- （8）删除四项评分均为0的记录，存为 cleaned_data_c8_N.csv；
- （9）将评分超过5分的无效值修正为5分，存为 cleaned_data_c9_N.csv；
- （10）识别并删除完全重复的记录（所有字段值都相同），存为 cleaned_data_c10_N.csv；

将完成清洗后的10个文件名截图粘贴至【提交结果.docx】。

（二）任务二：数据标注

1. 子任务一：酒店价格区间标注——经济型（≤300元/晚）、中端型（301-800元/晚）、高端型（801元/晚以上），新增“价格类型”列，写入hotel_price_level.csv（酒店名称、地区、房价、价格类型）；
2. 子任务二：酒店服务质量标注——服务质量得分=(服务评分+卫生评分)/2；优质服务（≥4.5分）、良好服务（4.0-4.4分）、标准服务（3.5-3.9分）、需改进服务（<3.5分），写入hotel_service_level.csv（酒店名称、服务评分、卫生评分、服务质量得分、服务等级）。

（三）任务三：数据统计

1. 子任务一：HDFS文件操作——创建目录结构/hotel/data和/hotel/backup；将hotel.csv上传至/hotel/data并验证；配置/hotel/data目录访问权限（所有者rwx、所属组r-x、其他用户r--）并验证；查看已上传文件前5行；将数据文件复制到备份目录并命名为hotel_backup.csv；
2. 子任务二：计算输入文件中的单词数——HDFS创建/user/hadoop/input；上传/var/log/dmesg；wordcount统计，结果存到/user/hadoop/output，查看前十行截图；
3. 子任务三：数独解算器——使用Hadoop默认sudoku示例完成数独题目，将解题结果截图粘贴。

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（使用pandas完成以下任务，结果在IDE控制台打印并截图）

- （1）分析各地区酒店的分布情况，统计每个地区的酒店数量，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，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，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。

2. 子任务二：数据可视化（使用pyecharts库绘制，图表截图粘贴至【提交结果.docx】）

- (1) 柱状图展示各个地区的酒店数量；
- (2) 条形图比较不同地区酒店的平均卫生评分和服务评分；
- (3) 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；
- (4) 散点图探索房价与评价数量之间的关系。

（二）任务二：业务分析

计算平均房价；确定最受欢迎的房间类型（根据评价数量）；识别评分在前25%的酒店共有的特征。根据分析结果撰写简短描述，提出两条针对酒店的营销策略建议，粘贴至【提交结果.docx】。

—— 本套赛题完 ——