

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四

赛项编号：ZZ052

整理时间：2026年8月

场景：旅游景区客流量预测与分析

旅游行业作为现代服务业的重要组成部分，其数据分析对提升游客体验、优化资源配置具有重要意义。通过收集和分析游客的行为数据（景点偏好、消费习惯、住宿选择等），可以为旅游服务提供者制定更精准的运营策略提供依据。为完成旅游景区客流量预测与分析工作，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	√		

服务器	master	slave1	slave2
YARN-NodeManager	√	√	√
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) 配置MySQL服务器的最大连接数为1000，提交配置命令；
- (2) 创建数据库 tourism 并设置默认字符集为utf8mb4，提交设置命令；
- (3) 创建一个名为 tourism_backup 的定时备份脚本，每天凌晨2点自动备份数据库tourism，提交脚本内容；
- (4) 优化MySQL查询性能，开启慢查询日志，设置超过3秒的查询被记录，提交配置命令；
- (5) 配置MySQL的binlog日志，设置过期时间为7天，提交配置命令。

3. 子任务三：数据表的创建及维护

在tourism数据库中创建景区信息表和游客流量表：

- (1) scenic_spot景区信息表：spot_id(int, 景区编号, 主键)、spot_name(varchar, 景区名称)、city(var char, 所在城市)、level(varchar, 景区等级)、type(varchar, 景区类型)、ticket_price(decimal, 门票价 格)、opening_hours(varchar, 开放时间)、max_capacity(int, 最大承载量)；
- (2) visitor_flow游客流量表：record_id(int, 记录ID, 主键)、spot_id(int, 景区编号, 外键)、visit_da te(date, 参观日期)、visitor_count(int, 游客数量)、peak_hour(time, 高峰时段)、weather(varchar, 天 气状况)、is_holiday(boolean, 是否节假日)；
- (3) 将两个SQL建表语句分别粘贴至【提交结果.docx】；
- (4) 导入数据后编写SQL：查询2024年1月每个景区的总游客人数（显示景区名称、所在城市、游客总数， 按游客总数降序排列）；统计每个景区在节假日和非节假日的平均游客数量（显示景区名称、等级、节假 日平均游客数、非节假日平均游客数，按景区编号排序）；找出所有景区中单日游客量超过该景区最大承 载量的情况（显示景区名称、参观日期、实际游客数量、最大承载量、超出人数，按超出人数降序排列） 。

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

有一份景区游客数据文件（景区名称、日期、游客来源地、游客年龄段、停留时长小时、消费金额元、满 意度评分、出行方式、天气状况等），已存入tourist_data.csv。使用pandas读取并打印前10行截图粘贴 。

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）删除天气状况为空或为零的记录，存为 `cleaned_data_c1_N.csv`；
 - （2）删除停留时长异常（超过24小时或小于0.1小时）的记录，存为 `cleaned_data_c2_N.csv`；
 - （3）将游客来源地为空的记录填充为“未知”，存为 `cleaned_data_c3_N.csv`；
 - （4）将满意度评分超出范围（0-5分）的记录删除，存为 `cleaned_data_c4_N.csv`；
 - （5）将消费金额明显异常（超过该景区平均消费金额5倍）的记录删除，存为 `cleaned_data_c5_N.csv`；
- 将该5个文件名截一张图粘贴至【提交结果.docx】。

（二）任务二：数据标注

1. 子任务一：游客价值分类标注——高价值游客（消费金额高于平均值1.5倍且停留时长超过4小时）、中价值游客（消费金额在平均值0.5-1.5倍之间且停留时长在2-4小时之间）、低价值游客（不满足上述条件），新增“游客价值类型”列，存入 `tourist_value_mark.csv`；
2. 子任务二：旅游旺季标注——旺季（月游客量超过年平均120%且月均消费超过年平均110%）、平季（月游客量在年平均80%-120%之间或月均消费在年平均90%-110%之间）、淡季（月游客量低于年平均80%且月均消费低于年平均90%），新增“季节类型”列，存入 `season_mark.csv`。

（三）任务三：数据统计

1. 子任务一：HDFS文件操作——新建目录 `/tourism_data`；赋予最高777权限；将本地旅游数据集上传至 `/tourism_data` 并验证上传成功，提交完整命令；
2. 子任务二：使用MapReduce进行数据分析——编写MapReduce程序统计不同景区的每日平均游客数量，结果存到HDFS的 `/tourism_analysis/daily_visitors` 目录，运行并查看执行结果，提交运行命令和结果前十行截图；
3. 子任务三：使用Hive进行数据分析——在Hive中创建外部表 `tourist_flow` 加载HDFS中的旅游数据；编写HQL查询：统计每个景区的月度游客总量、计算不同游客来源地的平均消费金额、统计不同天气状况下的游客总量；将查询结果导出到本地文件系统，提交导出命令和结果截图。

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）分析各景区的游客来源地分布，找出每个景区最主要的客源地（占比超过20%）；
- （2）计算不同年龄段游客的平均消费金额和停留时长，发现最具消费潜力的游客群体；
- （3）分析周末和工作日的游客量差异，计算周末游客量相对工作日的增长百分比；
- （4）统计不同出行方式的游客比例，并根据平均消费金额对出行方式进行排序；
- （5）分析满意度评分与停留时长的关系，计算相关系数。

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）玫瑰图展示各景区游客量分布（扇形区域半径表示游客量大小，颜色深浅代表游客量级别，支持鼠标悬停显示详细数据）；

（2）气泡图分析景区特征（横轴平均消费金额、纵轴日均游客量、气泡大小表示满意度评分、颜色表示景区等级，支持图例筛选和缩放）；

（3）堆叠面积图展示一年内各主要客源地游客量变化趋势（不同颜色代表不同来源地，显示月度数据点，支持区域高亮）；

（4）雷达图对比不同年龄段游客群体的消费特征（消费金额、停留时长、满意度评分、重游率、景点打卡数等维度，添加数据标签）。

图表必须包含完整标题、图例、坐标轴标签等要素。

（二）任务二：业务分析

（1）基础指标分析：客流指标（日均客流量=每日游客总数算术平均值、客流波动系数=标准差/平均值、节假日客流比=节假日日均客流/普通日均客流）；消费行为（人均消费=总消费金额/总游客人数、不同群体消费差异）；游客体验（满意度评分采用5分制，计算所有有效评分的加权平均值）。

（2）运营建议：从差异化票价策略（结合客流波动规律和游客消费能力设计分时段、分群体票价方案）、服务优化方案（针对满意度影响因素提出服务改进措施）、精准营销建议（基于高价值客群画像设计针对性营销策略）三方面提出建议。

将分析报告（不超过300字）提交至【提交结果.docx】，报告内容应包含数据分析结果和具体可行的运营建议。

—— 本套赛题完 ——