

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三

赛项编号：ZZ052

整理时间：2026年8月

场景：共享单车使用数据分析

共享单车作为一种新型城市短途出行解决方案，满足了人们「最后一公里」的出行需求。共享单车平台可以收集用户骑行轨迹、使用时段、车辆状态等多维度数据，通过分析可以识别热门区域和时段，预测车辆使用需求，优化调度策略，为城市交通规划提供参考。

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	√	√	√
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-libs-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）配置服务端MySQL数据库的远程连接，提交新增配置内容；
- （2）配置root用户允许任意IP连接，提交完整命令；
- （3）通过root用户登录MySQL，查看mysql库下所有表，提交命令及结果；
- （4）创建新用户 bike_admin，提交完整命令及结果；
- （5）创建数据库 bike_data，设置正确字符集，提交完整命令及结果；
- （6）授予新用户查询数据和插入数据的权限，提交完整命令及结果；
- （7）刷新权限，提交完整命令及结果。

3. 子任务三：数据表的创建及维护

在bike_data数据库中创建骑行记录表和单车状态表：

- （1）ride_records骑行记录表：id(int, 记录编号, 主键)、bike_id(varchar, 单车编号)、user_id(varchar, 用户ID)、start_time(datetime, 开始时间)、end_time(datetime, 结束时间)、start_location(varchar, 起始位置)、end_location(varchar, 结束位置)、ride_distance(double, 骑行距离km)、ride_duration(int, 骑行时长分钟)；
- （2）bike_status单车状态表：bike_id(varchar, 单车编号, 主键)、status(varchar, 车辆状态)、battery_level(int, 电量百分比)、last_maintain_time(datetime, 最后维护时间)、current_location(varchar, 当前位置)、total_mileage(double, 总里程km)；
- （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：在ride_records表中查询骑行时长超过60分钟的记录数量；在bike_status表中统计各个状态的单车数量；将两个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

有一份共享单车使用数据（骑行ID、单车编号、用户ID、开始时间、结束时间、起始位置、结束位置、骑行距离km、骑行时长分钟、天气状况、温度、周末标记），已存入bike_rides.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）删除骑行距离为空或为0的记录，存为 cleaned_data_c1_N.csv；

- (2) 删除骑行时长小于1分钟或大于24小时的异常记录，存为 `cleaned_data_c2_N.csv`;
- (3) 将天气状况中的空值替换为该时段最常见的天气状况，存为 `cleaned_data_c3_N.csv`;
- (4) 识别并删除重复的骑行记录，存为 `cleaned_data_c4_N.csv`;

将该4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：骑行距离分类标注——根据骑行距离分为短途（<2公里）、中途（2-5公里）、长途（>5公里），新增“距离分类”列，存入`ride_distance_mark.csv`（骑行ID、开始时间、骑行距离、距离分类）；
- 2. 子任务二：时段热度标注——高峰时段（工作日7:00-9:00和17:00-19:00）、平峰时段（工作日9:00-17:00和19:00-22:00）、低峰时段（工作日22:00-次日7:00及周末全天），新增“时段热度”列，存入`time_period_mark.csv`。

（三）任务三：数据统计

- 1. 子任务一：HDFS文件操作——新建目录/`bike_data`；赋予最高777权限；下载/`bike_data`到本地master的/tmp下，提交完整命令；
- 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：计算区域平均骑行时长——编写MapReduce程序，输入数据格式为“起始位置,结束位置,骑行时长(分钟)”，Map提取起始位置的区域名称和骑行时长，Reduce计算每个区域的平均骑行时长，输出到/`user/hadoop/avg_duration`，输出格式为“区域名称 平均骑行时长(分钟)”，将运行结果截图粘贴。

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) 按开始时间统计每个小时的平均骑行量，倒序排序展示前三个高峰时段；
- (2) 计算不同天气状况下的平均骑行时长和平均骑行距离；
- (3) 分析工作日和周末的骑行模式差异（平均骑行量、时长、距离）；
- (4) 统计每个区域的起始和结束骑行次数，识别最热门的上车区域和下车区域；
- (5) 分析温度对骑行量的影响，计算不同温度区间（0-10℃、10-20℃、20-30℃、30℃以上）的平均骑行量。

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) 时间热力图展示一周内各时段的骑行量分布（横轴星期、纵轴小时、颜色深浅表示骑行量）；
- (2) 折线图展示一天24小时内的骑行量变化趋势（工作日和周末两条折线对比）；
- (3) 区域流向图展示不同区域之间的骑行流量（箭头粗细表示流量，颜色表示平均骑行时长）；
- (4) 散点图分析气温与骑行量的关系（横轴温度、纵轴日骑行量、点大小表示平均骑行时长、颜色表示天气状况）。

（二）任务二：业务分析

（1）单车周转率：日均使用次数=单车每日骑行订单数平均值；日均使用时长=单车每日总骑行时长平均值（小时）；周转率=日均使用时长/24小时×100%；需分别计算工作日和周末的周转率。

（2）区域供需分析：供给量=该区域结束骑行的单车数量；需求量=该区域开始骑行的单车数量；供需平衡指数=(供给量-需求量)/需求量×100%；指数为正表示供大于求，为负表示供不应求，理想范围±20%之内。

（3）高峰期分析：对每个区域工作日早晚高峰（7:00-9:00，17:00-19:00），计算平均小时需求量、实际供给量、供需比=实际供给量/平均小时需求量；供需比低于1.2为供给不足，高于2.0为供给过剩。

根据上述分析结果撰写一份简短的分析报告（300字以内），包括当前运营情况的主要发现、存在的关键问题和挑战、至少三条具体的优化建议并说明预期效果，粘贴至【提交结果.docx】。

—— 本套赛题完 ——