

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一

赛项编号：ZZ052

整理时间：2026年8月

场景：房地产交易分析（二手房销售数据分析）

房地产市场正经历数字化转型。房地产大数据分析平台通过采集和处理海量的交易数据、用户行为数据和市场环境数据，全方位描绘市场格局。为完成二手房销售数据分析工作，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	√		

3. 子任务三：Zookeeper集群安装配置

- (1) 在master将 /opt/software 下的 apache-zookeeper-3.8.3-bin.tar.gz 解压到 /opt/module，提交解压命令；
- (2) 将解压目录更名为 zookeeper-3.8.3，提交命令；
- (3) 设置Zookeeper环境变量，提交新增内容；
- (4) 创建zoo.cfg配置三节点集群，dataDir设置为 /opt/module/zookeeper-3.8.3/data，提交配置内容；
- (5) master创建myid文件并设置内容为1；
- (6) 将Zookeeper环境变量文件及解压包拷贝到slave1、slave2，提交命令；
- (7) slave1修改myid为2、slave2修改myid为3，提交命令和截图；
- (8) 三节点分别启动Zookeeper，提交命令和截图；
- (9) 三节点分别查看Zookeeper状态，提交命令和截图；
- (10) master查看Java进程，提交命令和截图。

4. 子任务四：Kafka安装配置

- (1) 在master将 /opt/software 下 kafka_2.12-3.6.1.tgz 解压到 /opt/module，更名为kafka，提交解压命令；
- (2) 配置好Zookeeper（集群模式，master、slave1、slave2作为节点）；
- (3) 配置Kafka环境变量并生效，提交新增内容截图；
- (4) 使用 kafka-server-start.sh --version 查看版本，提交命令和结果截图；
- (5) 修改server.properties配置文件，分发Kafka文件到slave1、slave2，每个节点启动Kafka，提交启动命令和结果截图；
- (6) 创建Topic（名称为installtopic，分区数2，副本数2），提交创建命令和结果截图。

（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) 查看当前MySQL服务器状态和版本信息，提交命令和结果；
- (2) 创建 house_market 数据库，默认字符集utf8mb4，提交命令和结果；
- (3) 创建两个用户：house_admin（所有权限）和 house_viewer（只具有查询权限），提交命令和结果；
- (4) 为 house_market 数据库创建定时备份计划，每天凌晨2点自动备份，提交命令和结果；
- (5) 查看MySQL当前最大连接数和缓存大小配置，提交命令和结果；
- (6) 修改MySQL配置：最大连接数设为1000，查询缓存大小设为64MB，提交命令和结果；

- (7) 查看 house_market 数据库中所有表的存储引擎类型，提交命令和结果；
- (8) 创建 backup_user 用户，仅授予备份相关权限（BACKUP_ADMIN和RELOAD），提交命令和结果。

3. 子任务三：数据表的创建及维护

在house_market数据库中创建房源信息表(house_info)：

- (1) 根据题目提供的字段创建房源信息表（房源基本信息、价格、面积、户型、市区等字段）；
- (2) 将SQL建表语句粘贴至【提交结果.docx】；
- (3) 后续数据标注产出 area_popularity_mark.csv 区域热度标注文件。

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

使用pandas读取二手房销售数据CSV文件，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）。

基于二手房销售数据集进行清洗：删除空值记录、处理异常值、去重等，具体规则按题目要求，各清洗结果存为 cleaned_data_cN_N.csv。

（二）任务二：数据标注

基于数据集进行区域热度等标注，产出 area_popularity_mark.csv（打开后直接截图粘贴）。

（三）任务三：数据统计

- 1. 子任务一：HDFS文件操作——新建目录/file2_1；赋予最高777权限；下载/file2_1到本地master的/tmp下；
- 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统计并输出到/user/hadoop/output；查看结果前十行截图；
- 3. 子任务三：使用拟蒙特卡罗法估算Pi值——通过Hadoop Pi示例，使用16个映射（每个映射10,000个示例）估算pi值，提交结果。

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) 统计各市区的二手房销售数量，倒序展示；
- (2) 计算不同户型的平均售价；
- (3) 计算不同年份的二手房平均售价趋势；
- (4) 分析售价与面积之间的关系等。

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）柱状图展示各市区的二手房销售数量；
- （2）条形图比较不同户型的平均售价；
- （3）折线图展示不同年份的二手房平均售价趋势；
- （4）散点图探索售价与面积之间的关系。

（二）任务二：业务分析

计算平均售价；确定最受欢迎的户型（根据销售量）；识别销售量在前25%的市区共有的特征（售价范围、面积等）。根据分析结果撰写简短描述，提出至少两条针对二手房市场的营销策略建议，粘贴至【提交结果.docx】。

—— 本套赛题完 ——