# 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

# 3

# 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，为完成二手房销售数据分析工作，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。

3. 子任务三：Flume 安装配置

本任务需要使用 root 用户和 hadoop 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）从 master 中的 /opt/software 目录下将文件 apache-flume-1.11.0-bin.tar.gz 解压到 /opt/module 目录下，将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）把解压后的 apache-flume-1.11.0-bin 文件夹更名为 flume-1.11.0，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）使用 hadoop 用户完善 flume-env.sh 相关配置设置，配置 Flume 环境变量，并使环境变量生效，执行命令 flume-ng version 并将命令与结果截图粘贴至【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）在 MySQL 中创建一个新的数据库 namedb，并将创建命令复制粘贴至【提交结果.docx】中对应的任务序号下；

（2）将数据库的 root 用户的密码更改为一个新的强密码 12!@qwQW，并将更改密码的命令复制粘贴至【提交结果.docx】中对应的任务序号下；

（3）显示数据库中所有用户的权限信息，并将显示权限的命令及结果复制粘贴至【提交结果.docx】中对应的任务序号下；

（4）使用命令为数据库配置慢查询日志，记录查询超过 10 秒的 SQL 语句，并将配置命令及结果复制粘贴至【提交结果.docx】中对应的任务序号下；

（5）重启 MySQL 服务，并将重启命令及结果复制粘贴至【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 namedb 数据库中创建一个教师表（teacher）。教师表字段如下：

| 字段 | 类型 | 中文含义 |
| --- | --- | --- |
| id | int | 教师 ID |
| name | varchar | 姓名 |
| title | varchar | 职称 |
| academy | varchar | 学院 |

（2）根据以下数据字段在 namedb 数据库中创建一个选课表（enrollment）。选课表字段如下：

| 字段 | 类型 | 中文含义 |
| --- | --- | --- |
| id | int | 学号 |
| course_id | int | 课程 ID |
| score | int | 成绩 |

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）编写 SQL 查询，统计每门课程的选课人数，并将查询语句粘贴至【提交结果.docx】中。

（4）编写 SQL 查询，统计每门课程的平均成绩，并将查询语句粘贴至【提交结果.docx】中。

## 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

有一份二手房数据：市区、小区、户型、朝向、楼层、装修情况、电梯、面积(㎡)、价格(万元)、年份。

并且存入到 house_sales.csv 文件中，请使用 pandas 读取 house_sales.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

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）删除面积为空或为零的记录，并将结果存储为 cleaned_data_c1_N.csv，N 为删除的数据条数；

（2）删除“价格(万元)”为空或异常高（超过平均价格的 3 倍）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）对房型数据进行标准化，例如将不规范的房型描述（2 房间 2 卫）转换为标准格式（2 室 0 厅），并存储为 cleaned_data_c3_N.csv，N 为修改的数据条数；

（4）删除电梯字段为空白的记录，并将结果存储为 cleaned_data_c4_N.csv，N 为删除的数据条数；

（5）删除面积(㎡)小于 20 的记录，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：价格区间标注

使用 Python 编写脚本，根据房屋价格将房源分为“经济型”、“中档型”和“高端型”。具体的分类要求如下：

（1）**经济型**：价格低于该市区平均价格的 70%；

（2）**中档型**：价格在该市区平均价格的 70%至 130%之间；

（3）**高端型**：价格超过该市区平均价格的 130%；

在数据集中新增一列**“**价格区间**”**，根据上述标准对每个房源进行价格区间标注，存入 price_range_mark.csv 文件中。具体格式如下：

| 编号 | 小区 | 面积(㎡) | 价格（万元） | 价格区间 |
|---|---|---|---|---|
| 1 | 某某小区 | 90 | 300 | 中档型 |

将 price_range_mark.csv 打开后的直接截图（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：市区热门度标注

使用 Python 编写脚本，根据每个市区的房源数量标注市区的热门度。具体的分类要求如下：

（1）**高热门**：市区内房源数量高于全市平均数量的 120%；

（2）**中热门**：市区内房源数量在全市平均数量的 80%至 120%之间；

（3）**低热门**：市区内房源数量低于全市平均数量的 80%。

在数据集中新增一列**“**区域热门度**”**，根据上述标准对每个区域进行热门度标注，存入 area_popularity_mark.csv 文件中。具体格式如下：

| 编号 | 小区 | 市区 | 面积(㎡) | 价格(万元) | 区域热门度 |
|---|---|---|---|---|---|
| 1 | 育慧里一区 | 朝阳 | 52 | 343 | 高热门 |

将 area_popularity_mark.csv 打开后的直接截图（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：使用拟蒙特卡罗法估算 Pi 值

本任务需要使用 Hadoop 默认提供的 Pi 示例完成 Pi 值估算任务，具体要求如下：

（1）通过 Hadoop 提供的 Pi 示例，使用 16 个映射（每个映射 10,000 个示例）来估算 pi 值；

（2）将 Pi 值估算结果复制并粘贴至【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

数据分析对于理解房地产市场至关重要，可以揭示销售趋势、区域特点和市场需求。本任务中，我们将使用 Python 对二手房销售数据进行深入分析。参赛者需要使用 Python 的数据处理和分析库，如 Pandas 来完成以下任务：

（1）分析不同市区的二手房销售量，统计每个市区的房源销售数量，并进行倒序排序展示前三名；

（2）计算各户型的平均售价，并进行倒序排序展示前三种户型；

（3）计算每个市区的平均售价，进行倒序排序展示前三名；

（4）统计不同装修情况的房源数量，并进行正序排序展示；

（5）筛选出售价在平均售价以下且面积超过平均面积的房源，并统计这些房源的数量。

将这 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 Matplotlib 库来创建直观的图表，以揭示数据中的关键模式和趋势。具体要求如下：

（1）使用柱状图展示各市区的二手房销售数量，柱状图中的每个柱子代表一个市区，高度代表该市区的房源销售数量；

（2）创建条形图比较不同户型的平均售价，条形图中横轴表示平均售价，纵轴表示户型；

（3）使用折线图展示不同年份的二手房平均售价趋势；

（4）制作散点图探索售价与面积之间的关系，散点图中的横轴为售价，纵轴为面积，每个点代表一套房源。

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**

业务分析在二手房市场中至关重要，它可以帮助揭示客户需求、价格敏感度和市场发展趋势。在本任务中，我们将使用 Python 对二手房销售数据进行简单的业务分析，目的是识别市场的主要特征，并基于数据提出营销策略。

使用提供的二手房销售数据集，计算以下指标：

（1）**平均售价**：计算数据集中所有房源的平均售价。

（2）**最受欢迎的户型**：根据销售量，确定哪种户型最受欢迎。

（3）**高销量市区特征**：识别销售量在前 25%的市区共有的特征（如售价范围、面积等）。

根据上述分析结果，撰写一段简短的描述，提出至少两条针对二手房市场的营销策略建议。将内容复制粘贴至【提交结果.docx】中对应的任务序号下。