

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二

赛项编号：ZZ052

整理时间：2026年8月

场景：在线教育学习行为分析

在线教育平台的兴起为教育领域带来全新可能。通过收集和分析学习者在平台上的行为数据（课程选择、学习进度、作业完成情况、互动参与度等），可以更准确地把握学习者的需求和学习特点，建立个性化的学习档案，为每位学习者提供更有针对性的课程推荐和学习建议。

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-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）围绕在线教育场景创建数据库及用户（如 edu_admin 等），设置正确字符集，提交完整命令；
- （5）授予用户查询和插入数据的权限，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

围绕在线教育场景创建课程表、学习行为记录表等：

- （1）根据题目提供的数据库字段创建数据表（课程信息、学习进度、学习时长等字段）；
- （2）将SQL建表语句粘贴至【提交结果.docx】；
- （3）将提供的数据导入数据库并执行相应SQL查询操作。

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读取在线教育平台学习行为数据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）。

基于在线教育行为数据集进行清洗：删除空值、处理异常值、去重、填充等，各清洗结果存为 cleaned_data_cN_N.csv。

（二）任务二：数据标注

基于学习行为数据进行学习者价值分类、持续性分级等标注任务（如连续学习天数分级、学习时长等级等），存入对应标注文件并截图。

（三）任务三：数据统计

1. 子任务一：HDFS文件操作——新建目录并设777权限、上传/下载文件，提交完整命令；
2. 子任务二：计算输入文件中的单词数——wordcount统计dmesg单词数，查看结果前十行截图；
3. 子任务三：使用MapReduce/Hive对学习行为数据进行统计分析。

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）分析课程通过率分布；
- （2）分析学习时段分布规律；
- （3）分析持续性学习用户占比；
- （4）对比实战型课程与理论课的得分差异等。

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】）

使用pyecharts绘制学习行为相关图表（学习时段热力图、课程通过率柱状图、用户留存折线图 etc.）。

（二）任务二：业务分析

根据分析结果撰写业务报告，内容包括：主要发现（如实战型课程平均得分高于理论课19分等）、课程优化建议（为低通过率课程增加前置知识检测、拆分复杂模块为微课）、体验提升（开发移动端互动组件、建立21:00-22:00学习打卡激励）、留存策略（基于持续性分级设计阶梯式勋章奖励体系，连续学习>7天用户占29%）。

—— 本套赛题完 ——