2022年全国行业职业技能竞赛

第三届全国电信和互联网行业职业技能竞赛

计算机程序员（大数据分析）全国总决赛

竞赛技术手册

2022年12月

目录

一、 总决赛名称

（一）职工组

2022 年全国行业职业技能竞赛——第三届全国电信和互联网行业职业技能竞赛（大数据分析）全国总决赛

（二）学生组

2022 年全国行业职业技能竞赛——第五届全国大学生大数据技能竞赛全国总决赛

二、 总决赛方式及时间

（一）竞赛方式

本次总决赛为线上竞赛方式，不同组别分别登录竞赛通道。其中：

1.职工组竞赛链接：

（1）职工组：https://www.qingjiaoclass.com/matchdetail/gqprwvgpld

2.学生组竞赛链接：

（1）本科组：https://www.qingjiaoclass.com/matchdetail/boglaepjlx

（2）职教组：https://www.qingjiaoclass.com/matchdetail/aezlzdqzlb

3.竞赛开幕式：视频直播方式，竞赛前在群内发出直播地址。

（二）竞赛时间

1.开幕式时间：2022 年 12 月 20 日 13:30-14:00

2.总决赛时间：2022 年 12 月 20 日 14:00-18:00

3.监考录像审核：2022 年 12 月 21-23 日

三、 总决赛赛制安排

（一）竞赛内容

总决赛考核以大数据案例分析为主，竞赛总时间为 240 分钟，题型分为理论和实操两部分。理论题型占比 20%，知识点以 Python 技术和大数据技术为主，内容涵盖数据分析、数据建模、数据可视化等方向。实操占比 80%，以案例分析为主，共五个案例分别为使用 Spark 集群进行数据处理与分析、使用大数据组件开展数据分析处理、运用 Python 编写算法对文本类数据案例分析、运用 Python 编写算法对图像类数据案例分析、数据分析精准预测。

（二）竞赛评分机制

本次评分规则采用客观数据检测及过程检测综合评判办法，由专用的竞赛系统自动检测，并实时播报分数，竞赛结束后，由裁判组按统计每个模块得分，审核系统评分日志，监考录像无异议后，得出最终成绩。评分系统验证方式为代码或命令是否正确，结果是否正确，两者均正确情况下，系统自动给出得分，案例中每个任务先提交正确答案先得分，后提交会根据提交顺序按 1%进行递减计算分值，递减至 20%后不再递减，理论题型不递减。

理论题目在竞赛结束时由竞赛系统自动计算，题目答案正确得分，不正确不得分。

实操题目需要选手连接云主机按照要求进行操作。选手操作完成每个小任务时提交验证，当验证通过后，系统自动给出得分，每个任务先提交正确答案者得高分，后提交正确答案会根据提交顺序按 1%

进行递减，递减至20%后不再递减（例如：某实操题分值为100分，第一名提交得分为100，第二名得分为99，依次类推，第80名提交得分为20，第81名得分依然为20分），总成绩为所有完成任务得分的总和，如出现相同分数，排名按时间排序。

四、 竞赛规则

（一）竞赛秩序

（1）请所有选手自行做好竞赛准备，包括文件中第七部分中涉及的连接工具等；

（2）选手不得向竞赛平台做出任何网络攻防行为，一经发现取消参赛队竞赛资格，并由该参赛队承担一切损失；

（3）正式竞赛期间，经后台检测确定为非有效登录（单个账号有两个及以上IP登录）的团队和个人，视为违规不计入竞赛排名；

（4）竞赛期间请在竞赛系统中进行提问，提问方式见下方系统使用说明；

（5）选手对竞赛题目、过程提出的疑问，以竞赛组委会反馈的回复邮件为准；

（6）比赛过程中可以组内交流；严禁跨组交流，禁止参赛队伍之间分享任何解题思路，发现违规者一律取消参赛资格；学生组队员未在一起可用电话语音交流。

（7）竞赛选手不得使用任何社交软件（包括微信、钉钉、QQ、电子邮箱及其它任何带有社交功能的软件、通信平台），一经发现取消竞赛资格；

（8）未尽事宜将以竞赛组委会说明为准。

（二）取消竞赛或推迟竞赛的情况

（1）如发生重大事故、疫情或上级领导部门发布限制通知、公告， 可暂停或取消竞赛；

（2）因各种原因现场发生群体性事件，导致竞赛无法进行，可取消竞赛；

（3）竞赛平台因不可抗力导致无法正常进行竞赛，如停电、云端服务器大面积事故等，经裁判长决意后，可取消竞赛或推迟本次竞赛；

（4）竞赛延长原则，竞赛云端环境出现故障小于 30 分钟则不延长竞赛时间，超过 30 分钟不足 1 小时得延长 30 分钟。超过 1 小时不足 2 小时的，延长 1 小时。

（三）免责声明

本次竞赛涉及数据均为 2022 年全国行业职业技能竞赛——第三届全国电信和互联网行业职业技能竞赛计算机程序员（大数据分析）全国总决赛所需数据，仅用于对选手进行技术考核，竞赛中所涉及的信息均为脱敏、仿真数据，仅保留数据原始特征。如选手及其他人员需要对比赛设备（电脑、服务器等）的数据（文件、程序、信息等）进行处理，需要和组委会进行确认授权。

所涉及的竞赛信息，禁止随意传播、发布，对因使用本数据而导致的任何直接、间接、特殊、偶然或结果性损失概不承担责任。

（四）参赛注意事项

（1）参赛选手须自行准备电脑（配备摄像头），提前安装腾讯会议软件以及 Chrome 浏览器等竞赛要求终端软件（见竞赛链接附件要求），确保电脑可以连接互联网。

（2）参赛人员须实名使用电脑端腾讯会议软件进行登录（腾讯会议接入号：321-945-489），入会后全程打开摄像头，在竞赛开始前 10 分钟内，右手举身份证 30 秒，身份证正面对准电脑摄像头，远程核实身份。

1.职工组备注名为“单位简称+姓名"，其中：集团选拔赛选手以集团公司简称；省级选拔赛选手以省公司简称；组委会选拔赛选手以单位简称。如：移动集团+张三、辽宁电信+李四、清华大学-王五。

2.学生组备注名为“组别+团队名+姓名”，如：本科组+你说的都对+王五、职教组+data0i+李四。

（3）参赛选手另须自行准备带有摄像头智能手机一部，确保可以连接互联网，并安装手机版微信，扫描监考平台后，放置身后左 45 度位置，保证可以拍摄后背及电脑屏幕。

（4）竞赛选手竞赛过程中，须登录监考平台，开启录屏（录制桌面全屏模式），保证全程处于录制状态，竞赛过程中选手使用的电脑不得登录任何无关网页、社交账号，监考中或复核监 考录屏视频时，一经发现取消竞赛成绩。

五、 监考规则

竞赛期间采用全程监控功能，请每位参赛选手保证监控线路的正常开放。共计三路监控：

第一路监控为**腾讯会议**，开放本人摄像头。全程录制选手本人画面；

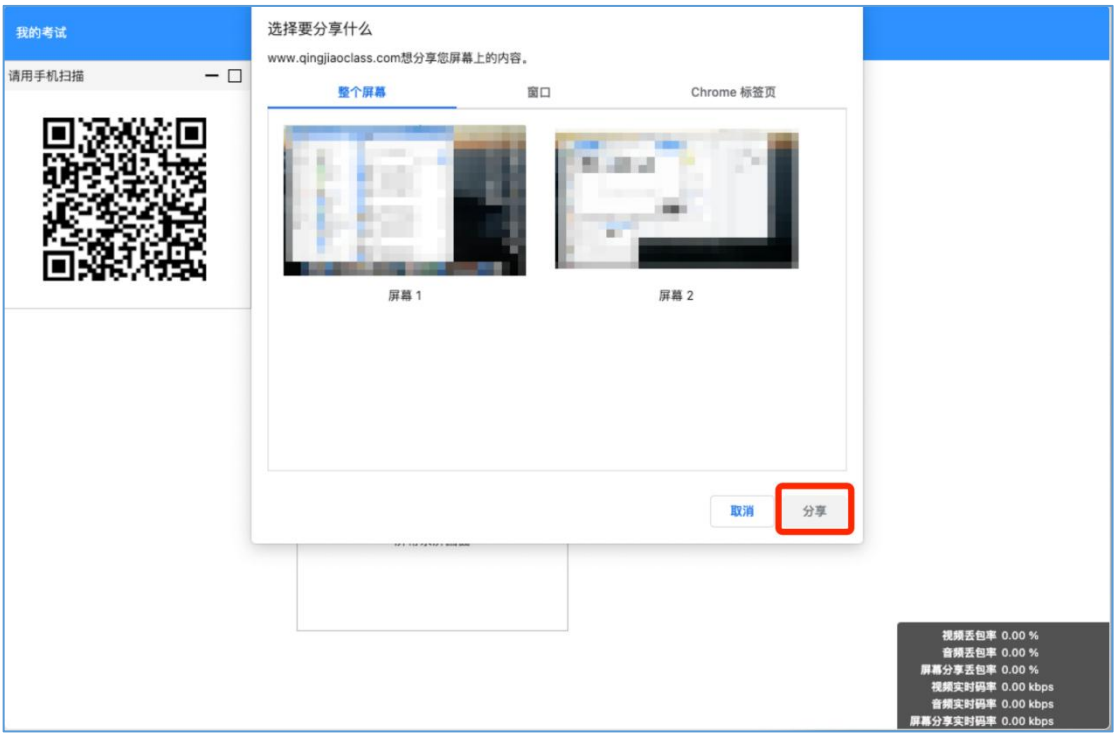
第二路监控为竞赛页面中“**监考—屏幕录制**”权限，全程录制本人在电脑上的操作行为，要求录制电脑全面桌面，如录制应用窗口，视为无效；

第三路为竞赛页面中“**监考—手机摄像头**”（副摄像头）权限，需要录制本人答题画面，使用微信扫描二维码即可，扫描监考平台后，放置身后左 45 度位置，保证可以拍摄后背及电脑屏幕。

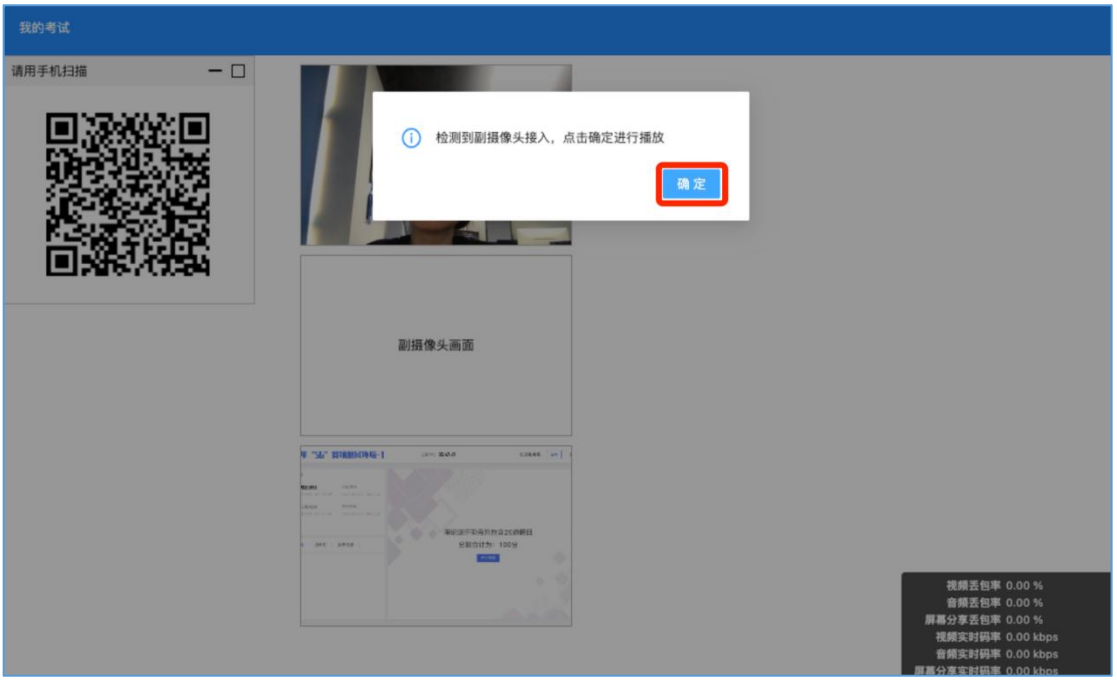
（一）监考功能使用方式

在进入正式竞赛页面后，在页面右上角“监控”按钮进，进入监考系统，进入监考系统后开放对应权限即可。

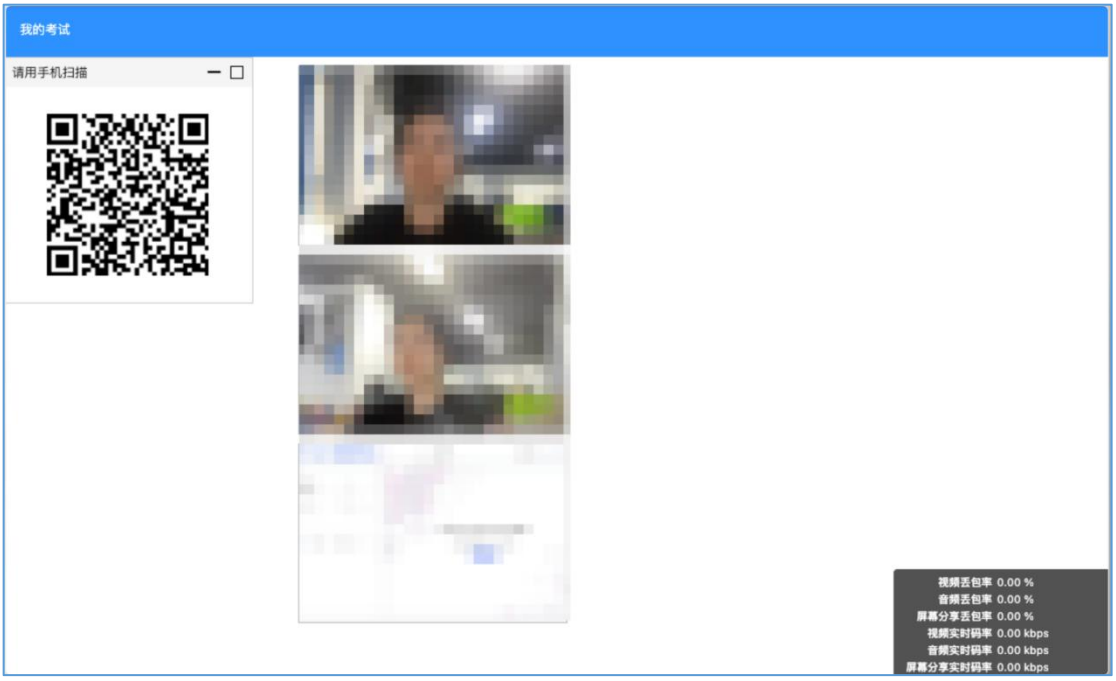
【图 1：进入“监考”示意图】



【图 2：开启屏幕录制权限示意图】



【图 3：开启副摄像头权限示意图】



【图 4：监控正常示意图】

（二）注意事项

（1）考试期间请使用**谷歌浏览器**进行答题，其他浏览器有可能部分功能不兼容；

（2）如未检测到监考画面可以尝试，请确认对应权限是否开启；

（3）监考页面关闭后，所有权限将实时退出。请选手在竞赛过程中避免关闭【图 4】所示的监考页面；

（4）副摄像头（手机端）页面退出监考后，摄像权限将实时退出，请选手在竞赛过程中避免关闭手机中的监考页面；

（5）竞赛过程中请保持【图 4】中监控画面为正常状态，监控开启后不得离开监控视频范围、不得故意遮挡摄像头；

（6）竞赛过程中保持在腾讯会议中开启摄像头。

六、 考核范围

考核模块	考核知识点与技能点	工具建议说明	职教组	本科组	职工组
理论	**题目一：理论模块** 模块包含数据分析、数据处理、大数据组件、算法分析与预测等方向内容。	无	√	√	√
大数据技术	**题目二：高校专业分析** 使用 Hadoop/Hive/Sqoop 等大数据技术对数据进行分析和处理，了解高校专业建设。 1.使用 Sqoop 进行数据迁移 2.读取 HDFS 数据 3.使用 Hadoop/Hive 进行统计分析 4.数据保存	本题目需要自备 Eclipse 等大数据开发工具链接云主机编写对应程序业务逻辑。也可自行使用其他开发工具	√	√	
	题目三：文本数据分析 结合 HanLP 工具，使用 Spark 技术对数据进行标准分词、关键词提取等数据操作。 1.结合 HanLP 进行中文/标准分词 2.使用 Spark 进行数据统计 3.数据保存		√		√

	题目四：网络设备数据分析 使用 Spark/Hbase 等大数据技术对数据进行监控分析，完成数据统计分析。 1.获取数据 2.使用 Spark 对设备进行统计分析 3.结果写入 Hbase		√	√	√
数据分析与挖掘	**题目五：高校专业分析** 使用 Python 技术对数据进行分析和处理，了解高校专业建设。 1.读取数据 2.数据清洗 3.数据筛选 4.数据分析 5.数据可视化 6.文件写入	本题目需要使用工具连接竞赛平台给出的 jupyter 环境，开启 jupyter notebook，打开对应网页即可进行操作。	√	√	
	题目六：文本数据处理 使用 Python 算法，对文本数据进行分词操作，查找文本相似度。 1.读取数据 2.数据清洗 3.文本清洗 4.jieba 分词器分词		√	√	√

	5.关键词提取 6.文本相似度 7.绘图/词云展示				
	题目七：图像数据处理 使用 Python 算法，对图像数据进行训练和预测。 1.读取数据 2.制作训练集 3.制作验证集 4.输出图像和标签 5.构建网络层 6.设置优化器 7.训练模型并保存 8.加载模型 9.预测结果			√	√
	题目八：对抗生成虚拟数据 使用对抗网络创建模型，生成虚拟数据，并对结果进行模型评估。 1.加载数据集 2.创建模型 3.损失函数 4.优化器				√

	5.训练模型 6.模型评估				

七、 环境及版本

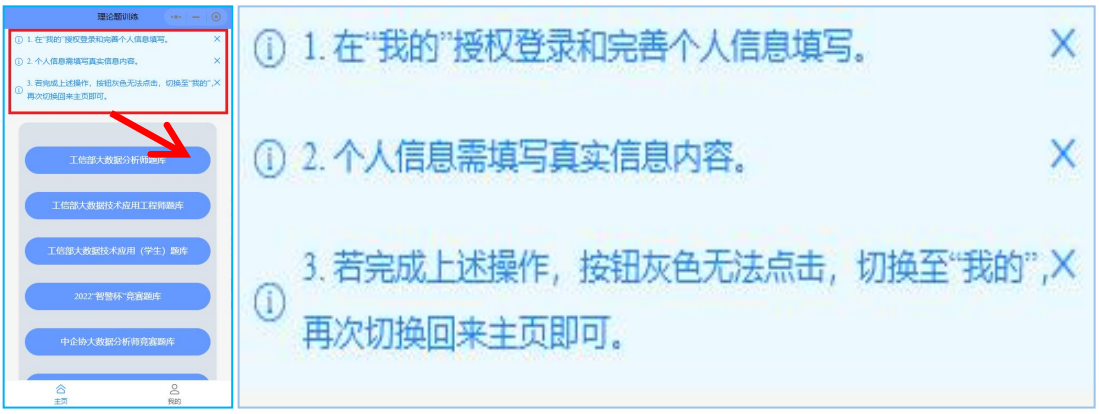
系统环境	CentOS Linux release 7.3.1611 (Core)
竞赛平台提供软件版本	Mysql 5.7.x（命令行形式）
	jdk-8u221-linux-x64.tar.gz
	hadoop-2.7.7.tar.gz
	spark-2.4.3-bin-hadoop2.7.tgz
	apache-hive-2.3.4-bin.tar.gz
	sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
	kafka_2.10-0.10.2.2.tgz
	hbase-1.6.0-bin.tar.gz
	apache-zookeeper-3.6.3-bin.tar.gz
	Python 3.x
选手根据个人习惯，本人电脑可提前准备软件	Chrome v94.0 或更高版本（此为推荐浏览器）
	Xshell 或 MobaXterm 等其他终端工具 如有安全考虑，可用电脑自带的 CMD 或 PowerShell
	eclipse 等其他开发工具

备注：以上仅为推荐软件，可根据使用习惯，选择相应软件工具，例如可自行安装图形化数据库管理工具代替终端数据库操作。

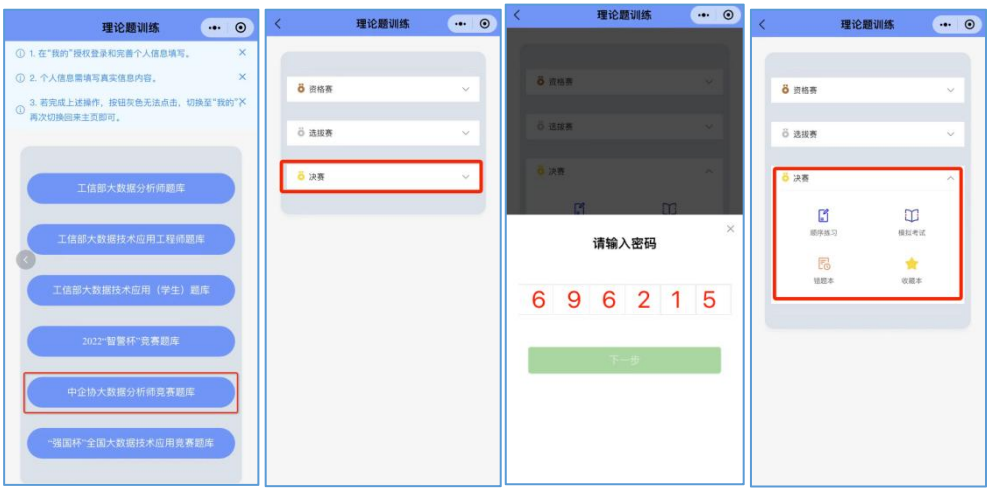
八、 理论题练习方式

1、微信搜索小程序“青椒小题库”或扫描下面二维码进入理论题学习小程序。

2.根据提示完成微信小程序授权即可使用小程序。



3.选择“中企协大数据分析师竞赛题库”→**决赛**→输入密码“696215”即可开始练习。



九、 竞赛系统使用说明

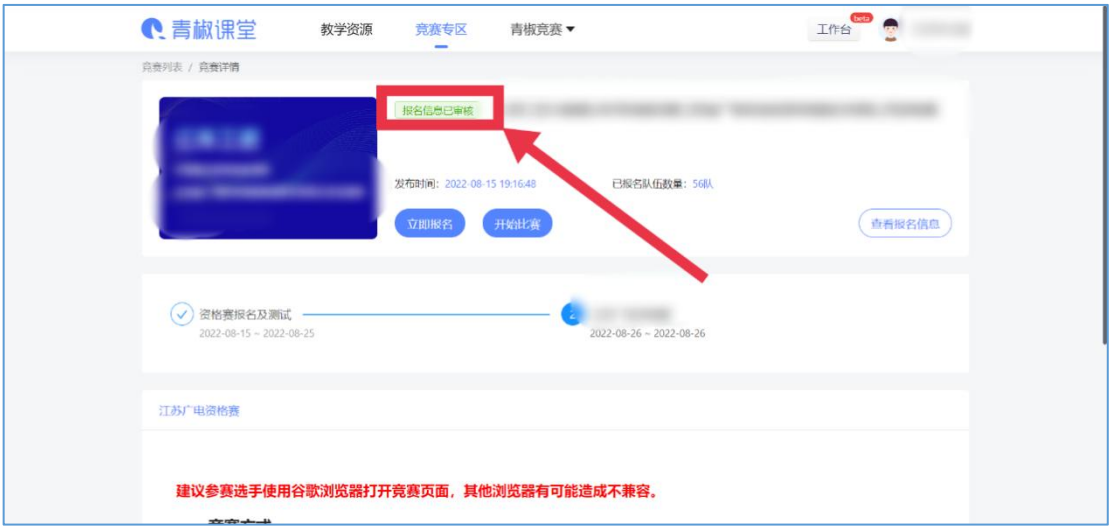
（一）登录青椒课堂

1.进入对应的总决赛正式竞赛链接，在竞赛时间内，点击“开始竞赛”即可进行竞赛。

注：本人账号建议设置成密码登录，牢记密码。

（二）系统使用说明

1.进入青椒课堂，竞赛专区，找到对应竞赛，通过青椒课堂竞赛页面的【开始比赛】进入竞赛页面；

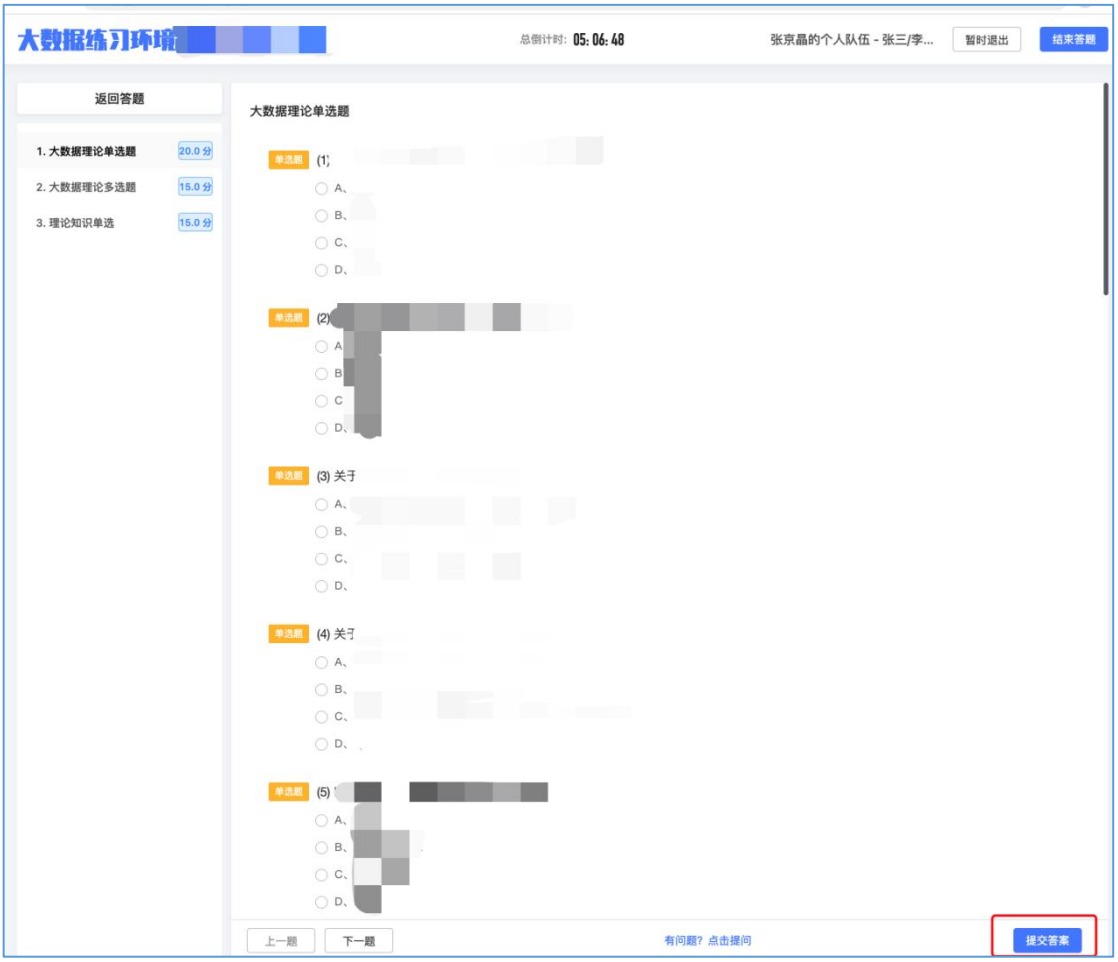


【开始比赛示意图】

2.进入竞赛页面后，若还未到开赛时间，将出现未开赛说明及倒计时；到达竞赛时间后，可以进入答题页面，点击开始答题即可；

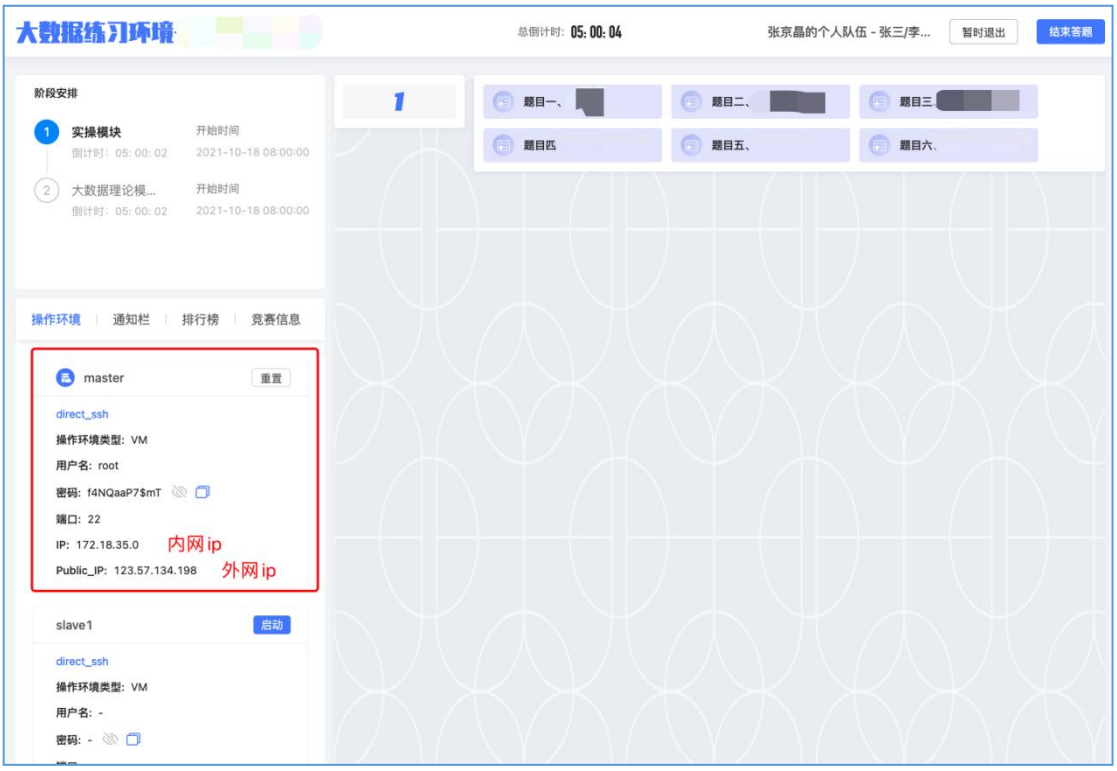
【答题页面示意图】

3.理论题目部分，选择答案后，点击“提交答案”，即可保存已选答案，在答题时间内，均可修改题目答案。完成理论题目后，点击返回答题，即可返回题目页面。（理论题目中每页题目都需要“提交答案”进行保存，在竞赛结束前均可进行修改。）



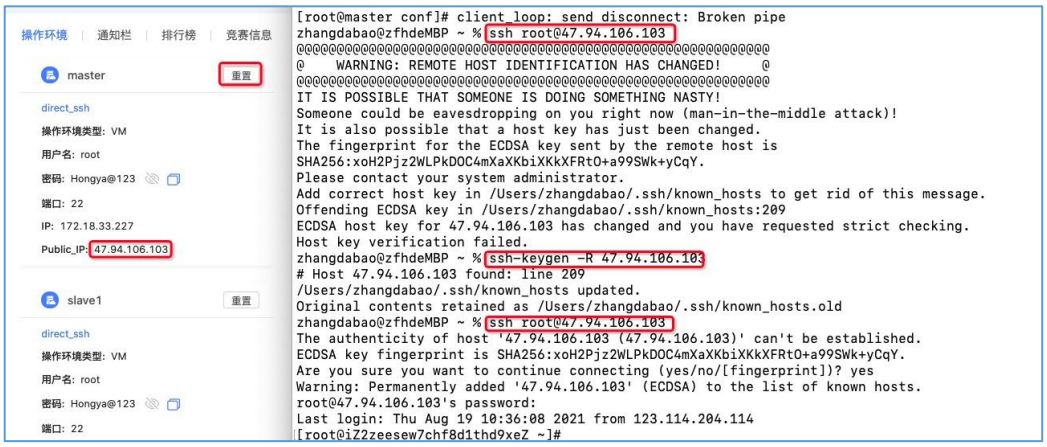
【理论答题页面示意图】

4.实操题目部分，需要启动虚拟机，虚拟机成功启动后，即可通过 Xshell 等连接工具,通过连接外网虚拟机 IP。连接虚拟机进行操作，集群连接使用内网 IP。



【IP 信息示意图】

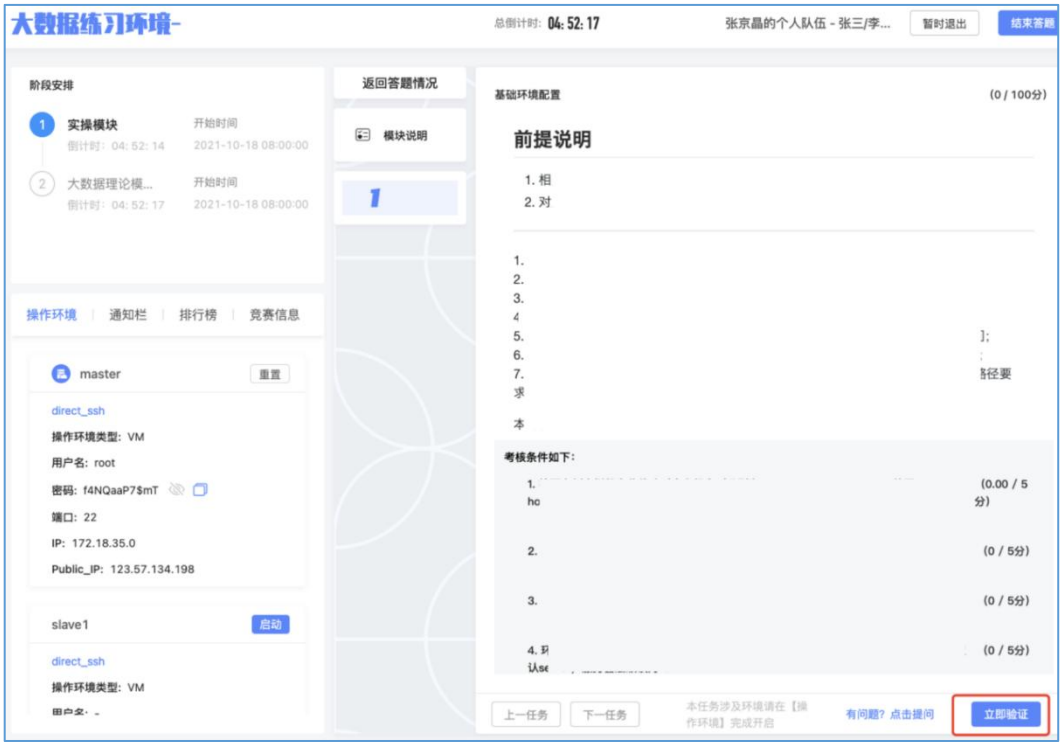
5.在竞赛过程重置属于初始化虚拟机操作，将清除之前所做的操作，对答题无任何帮助，建议不要轻易使用。重置不保证时效，谨慎操作。 如需重置虚拟机，申请重置后，请自行查看对应虚拟机 IP 地址。若 IP 地址无变化，选手重新操作时，本地电脑上已有上次连接认证储存在 known_hosts 文件中，再次连接会有报错，解决方式：删除对应 known_hosts 文件后重新连虚拟机或直接执行命令："ssh-keygen -R 虚拟机 IP"后重新连接虚拟机。



【重置后报错及解决示意图】

6.比赛验证为机器检测，全程无人为判分，保证比赛公平公正。赛题验证以结果为导向，不监控答题过程中的操作方式及方法。

注：选手完成每道实操题目后，需在此对应的步骤上点击**立即验证**，系统会判定当前步骤是否通过，如果通过可继续进行下一步骤，即状态变为绿色，否则需要继续修改直至完成。完成后每个条件的得分将实时同步，当步骤完成时会将用时总得分一起同步；



【立即验证示意图】

若出现验证不通过的情况，请在环境配置中自行查找原因进行修改。技术性问题比赛期间不做任何回复。

7.竞赛期间仅在平台内进行答疑，不在微信、QQ、钉钉进行任何回复，有疑问请在平台提问！！！

入口 1：通知栏

在通知栏【问题答疑】进行问题描述，也可截图复制进行提问，提问后技术回复的答案将在通知栏展现。（如果通知栏所有信息太多，可以在“只看我”按钮中选择开启，开启后，将只能看到自己的提问及针对问题的回答）

【提问入口-问题答疑示意图】

入口 2：页面右下角“有问题？点击提问”

在本入口进行的问题问，可针对题目进行提问，选择所提问的环节，并且可以使用截图复制，在问题框中附上截图说明问题。提问后

的技术回复，还是在通知栏中收取。

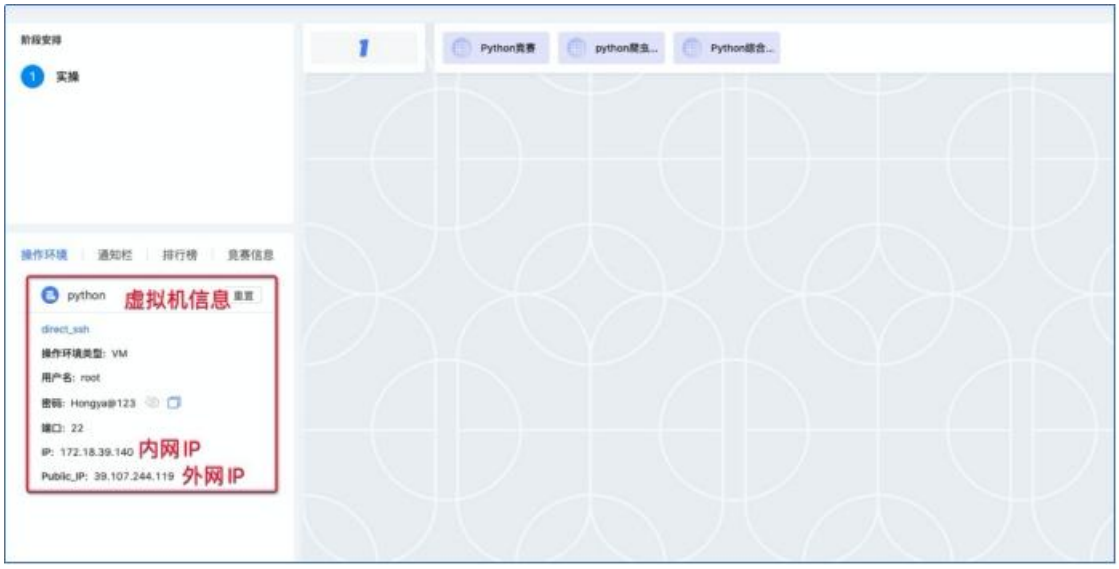
【提问入口–右下角‘点击提问’示意图】

8.暂时离开

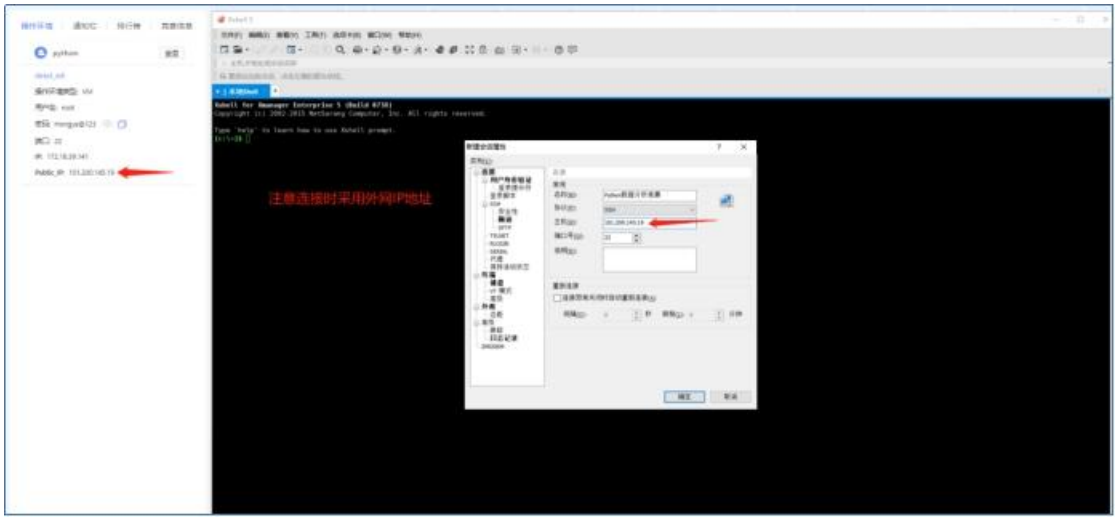
成功连入虚拟机且到达开赛时间，即可进入答题，竞赛页面右上角，【暂时离开】后还可以重新进入答题页面进行答题。

（三）竞赛平台连接说明-终端连接

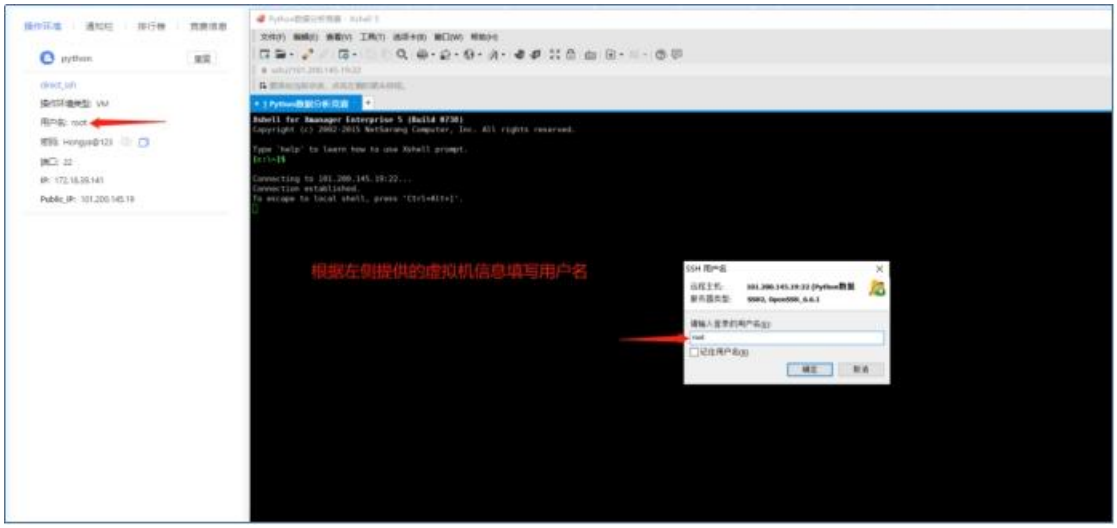
进入比赛后，可以在左侧栏看到“操作环境”选项，点击“开启”选项启动环境，启动后，可看到如下虚拟机信息，包括用户名、密码、端口、IP（内网 IP 和外网 IP），比赛中，需要选手连接虚拟机，从而对虚拟机进行操作，根据赛题要求完成验证得分。

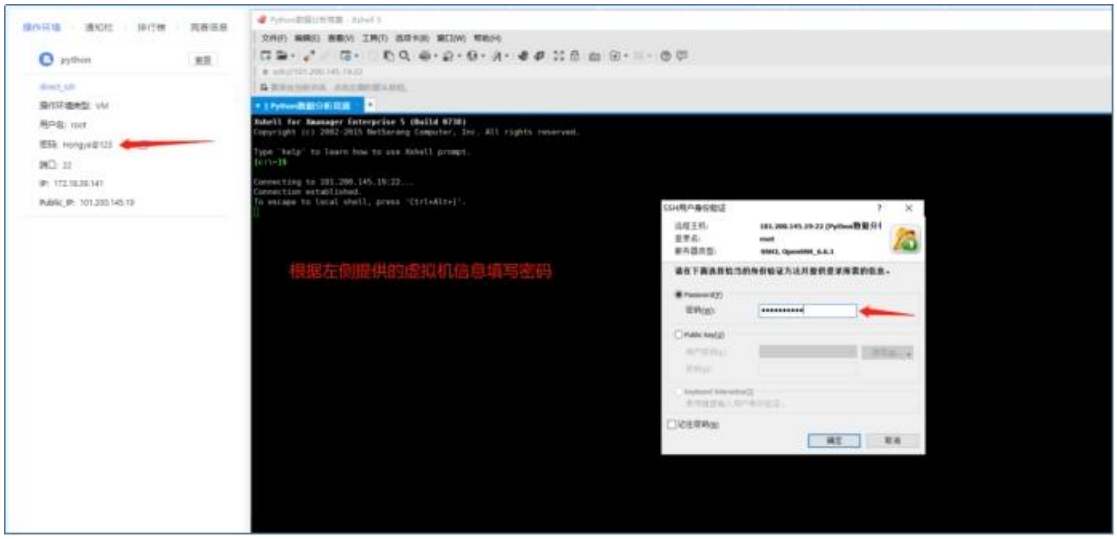


1.新建连接配置外网 IP 地址



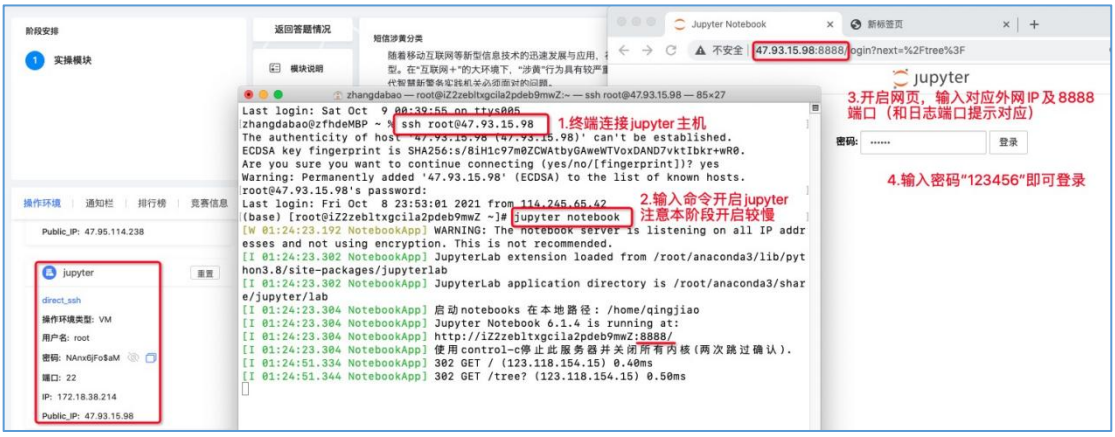
2.填写用户名和密码





（四）竞赛平台连接说明-jupyter 连接

1.环境中已经安装好 jupyter（anaconda 版本），开启虚拟机之后，终端连接虚拟机，执行“jupyter notebook”开启 jupyter,结合打印的端口号信息（如默认端口为 8888），直接在浏览器页面使用“IP:端口”的形式直接进行访问，注意以实际地址和端口为准。

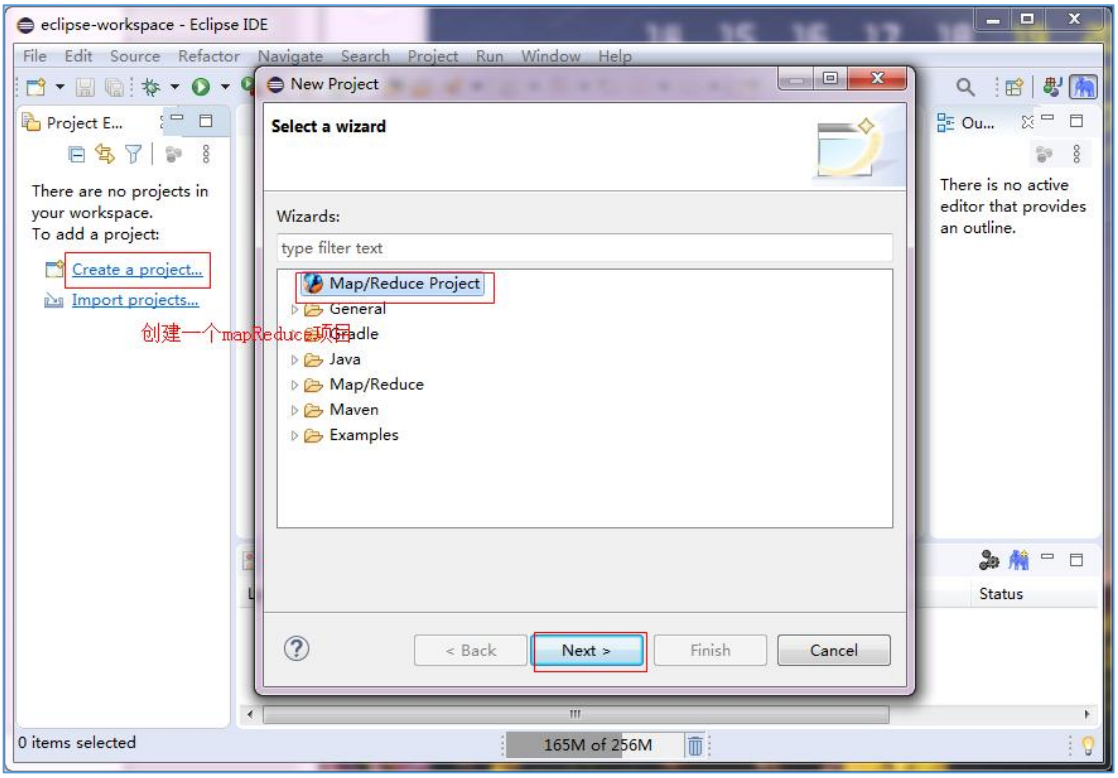


2.开启网页之后，就可以看到对应预置好的文件夹，进入即可看见相关文件和预置代码。

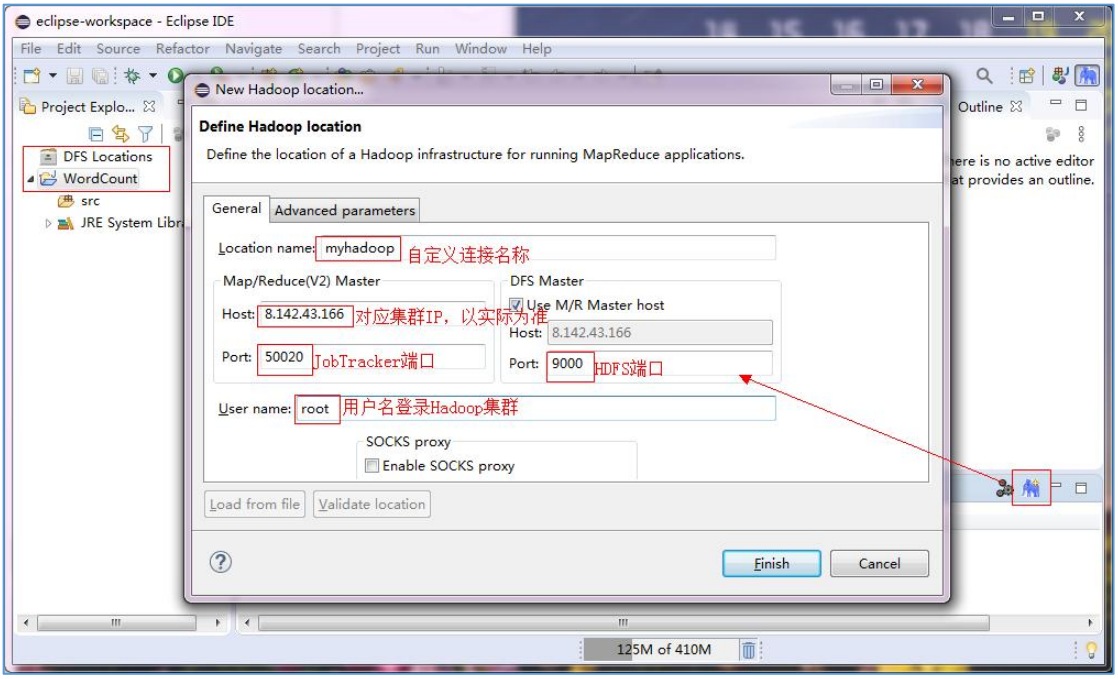
（五）竞赛平台连接说明-开发工具连接

1.本地安装需配置 Hadoop 环境，Eclipse 开发工具中增加 Map/Reduce 功能区。在 Eclipse 主界面的菜单中点击“window”-“Perspective”-“Open Perspective”-“Other”，弹出对话框中选择 Map/Reduce 选项，然后确定。

创建“Map/Reduce Project”MapReduce 工程。



2.对 Hadoop 集群的配置对话框进行编辑。



- Location name：命名新建的 Hadoop 连接，如 Hadoop Cluster。
- Map/Reduce(V2) Master：填写 Hadoop 集群的 ResourceManager 的 IP 和端口。

- DFS Master：填写 Hadoop 集群的 NameNode 的 IP 地址和连接端口。

3.编写程序进行数据分析，注意外域机器通信需要用外网 IP，未配置 hostname 访问会访问异常，可以在 Java api 客户端使用如下配置：

```
// 设置客户端访问 datanode 使用 hostname 来进行访问
conf.set("dfs.client.use.datanode.hostname", "true");
// 以实际 hotsname 为准
conf.set("fs.defaultFS", "hdfs://hostname:9000");
```