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

-,	总决赛名称	3
Ξ,	总决赛方式及时间	3
三、	总决赛赛制安排	4
四、	竞赛规则	5
五、	监考规则	7
六、	考核范围	11
七、	环境及版本	14
八、	理论题练习方式	15
九、	竞赛系统使用说明	16

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: <u>https://www.qingjiaoclass.com/matchdetail/gqprwvgpld</u>
 2.学生组竞赛链接:
- (1) 本科组: <u>https://www.gingjiaoclass.com/matchdetail/boglaepjlx</u>
- (2) 职教组: <u>https://www.qingjiaoclass.com/matchdetail/aezlzdqzlb</u>

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+dataOi+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。

2022年"56"赛项测试环境	总部计时: 02: 59: 49	晶」监考」 暂时退出
許段安排 1 理论模块 开始时间 倒计时:02:59:47 2022-09-07 18:08:50 2 実操模块 开始时间 例计时:02:59:47 2022-09-07 18:08:50		
操作环境 通知栏 竞赛信息	理论测评阶段共包含25道题目 分数合计为:100分 开始管题	

【图1:进入"监考"示意图】



【图2:开启屏幕录制权限示意图】

我的考试		
	① 检测到副摄像头接入,点击确定进行播放 确定	
	副損像头而面	
	The second secon	被捕去包率 0.00 % 音稱丟包率 0.00 % 厚幕分享玉包率 0.00 % 花網或封码率 0.00 kbps 音解或封码率 0.00 kbps 雷痛炎常母率 0.00 kbps 雷痛炎常母率 0.00 kbps 雷痛炎常母率 0.00 kbps 雷痛炎常母率 0.00 kbps

【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

老校		工目建议	职	本	职
与仪	考核知识点与技能点	工共建队	教	科	I
_ 模块 		西切	组	组	组
	题目一:理论模块				
理论	模块包含数据分析、数据处理、大数据组	无		\checkmark	
	件、算法分析与预测等方向内容。				
	题目二: 高校专业分析				
	使用 Hadoop/Hive/Sqoop 等大数据技术对	本题目需			
	数据进行分析和处理,了解高校专业建	要自备			
	设。	Eclipse 等	,		
	1.使用 Sqoop 进行数据迁移	大数据开	V		
上 **	2.读取 HDFS 数据	发工具链			
大致	3.使用 Hadoop/Hive 进行统计分析	接云主机			
が双	4.数据保存	编写对应			
	题目三: 文本数据分析	程序业务			
	结合 HanLP 工具,使用 Spark 技术对数据	逻辑。也可			
	进行标准分词、关键词提取等数据操作。	自行使用	/		,
	1.结合 HanLP 进行中文/标准分词	其他开发	V		V
	2.使用 Spark 进行数据统计	工具			
	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		V	V	\checkmark
	2.使用 Spark 对设备进行统计分析				
	3.结果写入 Hbase				
	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。	本题目需			
	1.读取数据	要使用工			
	2.数据清洗	具连接竞		\checkmark	
	3.数据筛选	赛平台给			
数据	4.数据分析	出的			
分析	5.数据可视化	jupyter 环			
与挖	6.文件写入	境,开启			
掘	题目六: 文本数据处理	jupyter			
	使用 Python 算法,对文本数据进行分词	notebook,			
	操作,查找文本相似度。	打开对应			
	1.读取数据	网页即可		\checkmark	\checkmark
	2.数据清洗	进行操作。			
	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		V	V
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			\checkmark
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)				
	Mysql 5.7.x(命令行形式)				
	jdk–8u221–Tinux–x64.tar.gz				
	hadoop–2.7.7.tar.gz				
	spark-2.4.3-bin-hadoop2.7.tgz				
竞赛平台提供软件版	apache-hive-2.3.4-bin.tar.gz				
本	sqoop–1.4.7.binhadoop–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				
4.14.1.2 年	Chrome v94.0 或更高版本(此为推荐浏览器)				
远于根据个人习惯,	Xshell 或 MobaXterm 等其他终端工具				
本 八 电 脑 り 徒 則 准 奋	如有安全考虑,可用电脑自带的 CMD 或 PowerShell				
私什	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。

理论数字结核 (***) (**) (*) ① 1 在 預約 (現在型素和応導不入氛围集等)、 × ① 2 个人氛围集等写真实信息内容。 × ② 1 名物成上过程件、按钮次点为法示击、切除室"我的",× ③ 1 名物成上过程件、按钮次点为法示击、切除室"我的",×	① 1.在"我的"授权登录和完善个人信息填写。	×
工作意大教派的折阅。	① 2. 个人信息需填写真实信息内容。	×
工总部人教部区本心用(学生) 题件 2022 智慧体 安吉勒体 中止协人教部分析特容高额件	3.若完成上述操作,按钮灰色无法点击,切换至"我 ① 再次切换回来主页即可。	韵",X

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。

理论题训练 💮	< 理论题训练	••• •• •	理论题训练	••• ••	< 理论题	UI\$\$ •• •
① 1. 在"我的"授权登录和完善个人信息填写。 ×						
① 2. 个人信息需填写真实信息内容。 X	Ö 资格赛		5 点档典		尚 资格赛	~
 3. 若完成上述操作,按钮灰色无法点击,切换至"我的"× 面次切换回来主页即可。 						
	Ö 选拔赛	~	5 追放费	× 1	◎ 选拔赛	v
工信部大数据分析师题库	<mark>ō</mark> 決赛	÷	决查	~	<mark>谈</mark> 决赛	~
			e 0	3	C	Ш
工信部大数据技术应用工程师题库			请输入密码	×	顺序练习	模拟考试
0					E	*
工信部大数据技术应用(学生)题库		e	6 9 6 2	15	48*	收 截本
2022"智管杯"竞赛题库						
中全协大数据分析师员寄题库			5- 5			
"强国杯"全国大数据技术应用竞赛题库						

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

北入青椒课堂, 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;

C .	青椒课堂 教学	资源 竞赛专区	青椒竞赛▼	工作台 📴 👮	
<u>育赛列</u> 提	1 / 寛賽洋情				
	-	报名信息已审核		a na maannaa da taal	
		及和时间: 2022-08-151 立即服名 并	9:1648 ERNEMAD	(2017) 200A (查看服名信息)	
(资格赛报名及测试 2022-08-15 ~ 2022-08-25		2022-08-26 ~ 20	22-08-26	
江水	广电资格赛				
i	建议参赛选手使用谷歌浏览	5器打开竞赛页面,其他 波	览器有可能造成不兼容。		

【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。)

大数据练习环	ф <mark>. П</mark>		总倒计时: 05: 06: 48	张京晶的个人队伍 - 张三/李	暂时退出 结束答题
返回答题		大数据理论单选题			
1. 大数据理论单选题	20.0 5	单选题 (1)			I
2. 大数据理论多选题	15.0 分	() A.			I
3. 理论知识单选	15.0 9) В.			I
		○ c.			
		00.			I 1
		单边题 (2)			
		O A .			
		00			
		O D.			
		^{●沈颢} (3) 关于			
		() A.			
) В.			
		⊖ c.			
		O D.			
		单选题 (4) 关于			
		○ A,			
		<u>о</u> в,			
		O D			
		(5)			
		<u>о в.</u>			
		⊖ c.			
) D.			
		上一题下一题		有问题? 点击提问	提交答案

【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。

大数据练习环境	总倒计时: 05:00:04	张京晶的个	人队伍 - 张三/李 暂时退出	结束答题
阶段安排 ① 实操模块 开始时间 倒计时:06:00:02 2021-10-18 08:00:00 ② 大数据理论模 开始时间		 第日二、 第日二、 第日五、 	 第目三、 第目三、 	
1911时: 05:00:02 2021-10-18:06:00:00 操作环境 通知栏 排行榜 竞赛信息				
 master 重置 direct_ssh 操作环境类型: VM 用户名: root 				
密码: f4NQaaP7\$mT				
slave1 語题 direct_ssh 操作环境类型: VM 用户名: - 密码: - 谈 ①				

【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。

操作环境 通知栏 排行榜 竞赛信息	[root@master conf]# client loop: send disconnect: Broken pipe zhangdaba02zfhdeMBP ~ % [ssh root@47.94.106.103]
(1) master 重置	0 WARNIG: REMOTE HOST IDENTIFICATION HAS CHANGED! 0 0 000000000000000000000000000000000
direct_ssh	IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY! Someone could be eavesdropping on you right now (man-in-the-middle attack)!
操作环境类型: VM	It is also possible that a host key has just been changed.
四白衣	The fingerprint for the ECDSA key sent by the remote host is
用广告: root	SHA256:xoH2Pjz2WLPkDOC4mXaXKbiXKkXFRtO+a99SWk+yCqY.
密码: Hongya@123 📎 🗍	Please contact your system administrator.
端口: 22	Add correct host key in /Users/zhangdabao/.ssh/known_hosts to get rid of this message. Offending ECDSA key in /Users/zhangdabao/.ssh/known_hosts:209
IP: 172.18.33.227	ECDSA host key for 47.94.106.103 has changed and you have requested strict checking.
Public IP: 47 94 106 103	Host key verification failed.
	zhangdabao@zfhdeMBP ~ % [ssh-keygen -R 47.94.106.103]
	# Host 47.94.106.103 found: line 209
Clave1 € €	/Users/zhangdabao/.ssh/known_hosts updated.
Slave I E	Original contents retained as /Users/zhangdabao/.ssh/known_hosts.old
direct cch	zhangdabao@zfhdeMBP ~ % [ssh root@47.94.106.103]
direct_ssit	The authenticity of host '47.94.106.103 (47.94.106.103)' can't be established.
操作环境类型: VM	ECDSA key fingerprint is SHA256:xoH2Pjz2WLPkDUC4mXaXKbiXKkXFRt0+a99SWk+yCqY.
用户名: root	Are you sure you want to continue connecting (yes/no/[fingerprint])? yes Warning: Permanently added '47.94.106.103' (ECDSA) to the list of known hosts.
密码: Hongya@123 🛞 🗇	root@47.94.106.103's password:
端口: 22	Last login: Thu Aug 19 10:36:08 2021 from 123.114.204.114 [[root@iZ2zeesew7chf8d1thd9xeZ ~]#

【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;

大数据练习环境-		总倒计时: 04: 52: 17	张京晶的个人队伍 - 张三/李	暂时退出 结束答题
阶段安排	返回答题情况	基础环境配置		(0 / 100分)
1 实操模块 开始时间 街计时: 04:52:14 2021-10-18 08:00:00	€ 模块说明	前提说明		
2 大数据理论模 开始时间 倒计时:04:52:17 2021-10-18 08:00:00	1	1. 相 2. 对		
操作环境 通知栏 排行榜 竞赛信息		1. 2. 3. 4 5. 6. 7. 求]; ; 若径要
操作环境类型:VM 用户名:root		考核条件如下:		
密码: f4NQaaP7\$mT 🐼 🗇 编口: 22		1		(0.00 / 5 分)
IP: 172.18.35.0 Public_IP: 123.57.134.198		2.		(0 / 5分)
slave1 启动		3.		(0 / 5分)
direct_ssh		4. 9 ilas		(0 / 5分)
操作环境类型: VM 田户오·-		上一任务下一任务	本任务涉及环境请在【操 有问题?点; 在环境】幸成开房	由提问 立即验证

【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)

大数据练习环境-		总倒计时: 04: 48: 14	张京晶的个人队伍 - 张三/李 暂时	退出 结束答题
阶段安排	返回答题情况	基础环境配置		(0 / 100分)
实操模块 开始时间 倒计时:04:48:11 2021-10-18 08:00:00	€ 模块说明	前提说明		
2 大数据理论模 开始时间 倒计时: 04:48:15 2021-10-18 08:00:00	1	1.1 2.3		
<u>ا</u> إذا	题答疑		×	
操作环境 通知栏 排行榜 竞赛信	问题描述:			ę
		nostnamectiap 약)	取消 发送 r、slave1、slave2,使用	(0.00 / 5 分)
		2.		(0 / 5分)
		3.		(0 / 5分)
		4. iλ,		(0 / 5分)
只看我: 🟦		上一任务	本任务涉及环境请在【操 有问题? 点击提问 作环境】完成开启	立即验证

【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。

		返回答题情况	基础环境配置			(0 / 100
1 实操模块 开始时间 图计时: 04:45:15 2021-10) 0-18 08:00:00	🖅 模块说明	前提说明			
2 大数据理论模 开始时间 倒计时: 04:45:19 2021-10) 0-18 08:00:00	1	1. # 2. 3			
	问是	题答疑		×		
作环境 通知栏 排行榜	• 竞赛信	问题条件:		v]		
		问题描述:		1		
				1		
				•		
				。 取2月 <u>发送</u> r、s	lave1、slave2,使用	(0.00 / 5 分)
	L		2.	。 取用 发送 ^{r、 s}	iave1、siave2、使用	(0.00 / 5 分) (0 / 5分)
	ŀ		2. 3.	取消 <u>发送</u> r、s	lave1、slave2、使用	(0.00 / 5 分) (0 / 5分) (0 / 5分)
	ł	(2. 3. 4.	取消 <u>发送</u> r. s	lave1、slave2,使用	(0.00 / 5 分) (0 / 5分) (0 / 5分) (0 / 5分)

【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。

阶段变得	1	Python素要	python展主	Python##		
AN AN						
操作环境 通知栏 排行物 竞赛信息						
python 虚拟机信息 ^{单五} direc_ssh						
用作名: root 图明: Hongya@123 ③ 〇 順曰: 22						
P: 172.18.39.140 内网P Public_IP: 39.107.244.119 外网IP						

1.新建连接配置外网 IP 地址

HHRE BOC NON REEL	a Net1	1993年11年11年11年11年11年11年11年11年11年11年11年11年11
0	17 - 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	
and at shift and		
Bit magnetist ()	Rabell for Amonger Indexprise 5 (Halls 478) Chargeright 11:1002 2013 McClaring Computer, Soc. ACL rights inserient: Type "help" in laws have to an Adult product.	
RC II R ULARA	(0)/03(1) (0)/03(1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(7 ×
Add P III.2010 II		

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。

段安排	返回答题情况	姬信涉黄分类	Jupyter Notebook	×	3 新标签页	× +
2 实操模块	图 模块说明	随着移动互联网等新型信息技术的迅速发展 型。在"互联网+"的大环境下,"涉黄"行为具 代和智新警务实践机关心须面对的问题。	500用,1 ← → C ▲ 不安至 47.93.15.	98:8888/0		/ter
	🖲 😑 🔵 ① zh	angdabao — root@iZ2zebltxgcila2pdeb9mwZ:~	- ssh root@47.93.15.98 - 85×27	3	开启网页, 新	入对应外网IP及888
	Last login: Sat O	ct 9 00:39:55 on ttys005	4 AW 1 + 4 + - + + + 1	- 1	口(和日志如	
	The authenticity	BP ~ % ssh root@47.93.15.98	於师注按 JUPyter 土机	-		
	ECDSA key fingerp	rint is SHA256:s/8iH1c97m0ZCWAtby	GAweWTVoxDAND7vktIbkr+wR0.	密约	c	登录
	Are you sure you	want to continue connecting (yes)	no/[fingerprint])? yes			
	Warning: Permanen	tly added '47.93.15.98' (ECDSA) 1	to the list of known hosts.		4.输入密	码"123456"即可登录
作环境 通知栏 排行榜 赛赛信息	Last login: Fri 0	s password: ct 8 23:53:01 2021 from 114.245.	65.42 2.输入命令开启jupyter	1		
	(base) [root@iZ2z	ebltxgcila2pdeb9mwZ ~]# jupyter n	otebook 注意本阶段开启较慢	1		
Public_IP: 47.95.114.238	[W 01:24:23.192 N	otebookApp] WARNING: The notebook	server is listening on all IP add	t		
	esses and not usi	ng encryption. This is not recomm	nended.			
	hon3.8/site-nacka	nes/junyterlab	ioaded ifom /foot/anaconda3/iib/py			
Jupyter III	[I 01:24:23.302 N	otebookApp] JupyterLab applicatio	on directory is /root/anaconda3/shar	5		
direct_ssh	e/jupyter/lab					
操作环境类型: VM	[I 01:24:23.304 N	otebookApp] 启动notebooks 在本地)	各径: /home/qingjiao			
田户名·roat	[I 01:24:23.304 N	otebookApp] http://iZ2zebltxgcila	2pdeb9mwZ:8888/			
	[I 01:24:23.304 N	otebookApp]使用control-c停止此服	务器井关闭所有内核(两次跳过确认).			
ate: NANXOFOSAM ()	[I 01:24:51.334 N	otebookApp] 302 GET / (123.118.15	64.15) 0.40ms			
第日: 22	[I 01:24:51.344 N	otebookApp] 302 GET /tree? (123.)	18.154.15) 0.50ms			
IP: 172.18.38.214	U					
Public IP: 47.93.15.98						

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。

💭 jupyter	Quit 注销
文件 运行 集群	
选择操作对象.	上传 新建一 4
	名字 最后修改 File size
GBDT	25 分钟前
	3 小时前
□ □ 随机森林	40 分钟前

(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/Re duce 功能区。在 Eclipse 主界面的菜单中点击"window" - "Perspe ctive" - "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。

😂 eclipse-workspace - Eclipse II	DE	×
File Edit Source Refactor	Navigate Search Project Run Window Help	
📑 • 🖬 🗟 💠 • O • 9	New Project	Q 🗄 😰 🛛 🕷
Project E 🗧 🗖	Select a wizard	╊ Ou 🛛 🗖 🗖
□ 🕏 🍸 🔛 🕴		6 8
There are no projects in		There is no active
your workspace.	Wizards:	an outline.
	type filter text	
<u>Create a project</u>	Map/Reduce Project	
创建	▷ 🥭 General	
Rûxe i mab	> 🦻 Java	
	D 🥭 Map/Reduce	
	⊳ 🥭 Maven	
	Examples	
		ם - 🕌 🍓
1		Status
	Cancel	
	III	۰.
0 items selected	165M of 256M	1 🖓

2.对 Hadoop 集群的配置对话框进行编辑。

🖨 eclipse-workspace - Eclip	se IDE	– 🗆 X
File Edit Source Refact	tor Navigate Search Project Run Window Help	
1 - 🔛 💿 🕸 - Or		n q 🖻 🕏 🦣
Project Explo 🛛 🥤		Outline 🛛 🗖 🗖
E 🕏 7 s	Define Hadoop location	1
DFS Locations	Define the location of a Hadoop infrastructure for running MapReduce applications.	tere is no active editor
WordCount		at provides an outline.
JRE System Libri	General Advanced parameters	
_	Location name: myhadoop 百宁心连接复数	
	Man/Reduce(V2) Master	
	WUse M/R Master host	
	Host: 8.142.43.166 对应集群IP, 以实际为准 Host: 8.142.43.166	
	Port: 50020 JobTrackergh	
	User name: root 用户名登录Hadoop集群	
	SOCKS proxy	
	Enable SOCKS proxy	
	Load from file	
	() Einish Cancel	
U		۳ L
*		•
	125M of 410M	0

- 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");