全国职业院校技能大赛
赛项规程

赛项名称：大数据应用与服务

英文名称：Big data applications and services

赛项组别：中等职业教育

赛项编号：ZZ052

一、赛项信息

<table>
<tr><td colspan="4">赛项类别</td></tr>
<tr><td colspan="4">☐每年赛　☑ 隔年赛（☐单数年/☑ 双数年）</td></tr>
<tr><td colspan="4">赛项组别</td></tr>
<tr><td colspan="4">☑ 中等职业教育　☐高等职业教育</td></tr>
<tr><td colspan="4">☑ 学生赛（☐个人/☑ 团体）　☐教师赛（试点）　☐师生同赛（试点）</td></tr>
<tr><td colspan="4">涉及专业大类、专业类、专业及核心课程</td></tr>
<tr><td>专业大类</td><td>专业类</td><td>专业名称</td><td>核心课程
（对应每个专业，明确涉及的专业核心课程）</td></tr>
<tr><td rowspan="10">71 电子与信息大类</td><td rowspan="10">7102 计算机类</td><td rowspan="4">710205 大数据技术应用</td><td>大数据运维技术</td></tr>
<tr><td>大数据分析与挖掘技术</td></tr>
<tr><td>数据库应用技术</td></tr>
<tr><td>数据可视化技术</td></tr>
<tr><td rowspan="2">710201 计算机应用</td><td>计算机网络基础</td></tr>
<tr><td>数据库应用技术</td></tr>
<tr><td rowspan="3">710203 软件与信息服务</td><td>操作系统基础</td></tr>
<tr><td>数据库应用技术</td></tr>
<tr><td>程序设计基础</td></tr>
<tr><td colspan="4">对接产业行业、对应岗位（群）及核心能力</td></tr>
<tr><td>产业行业</td><td colspan="2">岗位（群）</td><td>核心能力
（对应每个岗位（群），明确核心能力要求）</td></tr>
<tr><td rowspan="8">战略性新兴产业-新一代信息技术</td><td colspan="2" rowspan="3">大数据平台运维</td><td>大数据平台搭建、部署与管理</td></tr>
<tr><td>Linux 操作系统管理与维护</td></tr>
<tr><td>数据库建库、建表与 SQL 数据处理</td></tr>
<tr><td colspan="2" rowspan="2">数据获取与清洗</td><td>对数据进行采集、加载和存储</td></tr>
<tr><td>数据标准化、数据清理、数据转换、数据验证</td></tr>
<tr><td colspan="2" rowspan="2">数据分析与可视化</td><td>数据的统计汇总、分区操作、分组操作</td></tr>
<tr><td>数据可视化效果的开发</td></tr>
<tr><td colspan="2" rowspan="2">计算机软硬件操作</td><td>熟练操作计算机和应用办公软件</td></tr>
<tr><td>常用软件工具的使用</td></tr>
</table>

二、竞赛目标

习近平总书记指出“大数据是信息化发展的新阶段”， 国家“十四五”规划也提出“打造数字经济新优势，充分发挥海量数据和丰富应用场景优势，促进数字技术与实体经济深度融合，赋能传统产业转型升级，催生新产业新业态新模式”，大数据成为推动社会发展的强大动力。本赛项旨在落实国家建设数字中国战略，大力推进大数据技术及相关专业的技术技能型人才培养，全面提升相关专业学生的综合能力，展现选手团队合作、工匠精神等职业素养，赋能经济社会高质量发展。

本赛项内容围绕大数据相关产业岗位的实际技能要求进行设计，重点考察参赛选手在大数据、数据库等方面的知识，以及大数据项目分析及实施、数据获取、加工和处理等方面的能力，还包括职业道德、工作态度、人际交往、团队合作、工匠精神等方面的素养。

三、竞赛内容

本赛项涉及大数据行业的典型工作场景，包括大数据平台搭建、数据库运行维护、数据清洗、数据标注、数据分析、数据可视化和业务分析等工作任务，考查的主要技术技能如下:

1）大数据平台搭建: 安装 Hadoop 全分布式平台，安装 Hadoop 平台相关的常用组件,包括但不限于ZooKeeper、Flume、Kafka、Spark、Flink、Redis、HBase 等，验证Hadoop 平台和相关组件的可用性。

2）数据库配置维护: 基于 MySQL 数据库进行建库建表，运用基本的 SQL 语句完成数据的增删改查等操作。

3）数据获取与清洗: 读取 CSV 数据源，对指定字段进行有效性检查，正确处理无效值和异常值，对数据进行一致性检查，对数据进行清洗和转换。

4）数据统计: 编写 Java MapReduce 程序，并将程序打包部署到Hadoop 平台上运行，对数据进行统计汇总、分区分组和排序等操作。使用 HDFS 上传和下载文件。

5）数据标注: 使用 Python 程序对数据进行分类标注，将标注后的数据保存到指定位置。

6）数据可视化: 使用 Web 技术或 Python 可视化技术对数据进行呈现，包括但不限于柱状图、折线图、玫瑰图、气泡图、饼状图、条形图、雷达图、散点图等效果。

7）业务分析: 能够理解业务场景，对业务数据进行分析，编写分析报告。

表 1 赛项内容与分值

模块		主要内容	比赛时长	分值
模块一：平台搭建与运维	任务一：大数据平台搭建	Hadoop 平台的安装部署和常用组件的安装部署。	180 分钟	10
	任务二：数据库配置维护	使用 MySQL 数据库建库建表，运用基本的 SQL 语言完成数据的增删改查等操作。	180 分钟	20
模块二：数据获取与处理	任务一：数据获取与清洗	对 CSV 数据文件进行加载、清洗和转换等操作，识别和处理无效值，检查数据的一致性，将清洗后的数据保存到指定位置。	120 分钟	10
	任务二：数据标注	使用 Python 语言对数据进行分类标注。	120 分钟	10
	任务三：数据统计	基于 Hadoop 平台进行编译、打包、部署和执行程序，完成数据的统计工作。	120 分钟	15
模块三：业务分析与可视化	任务一：数据可视化	使用 Web 前端框架或者 python 可视化库对数据进行可视化展示。	180 分钟	20/15
	任务二：业务分析	报表分析，对大数据项目的业务场景和数据进行分析，撰写报告。	180 分钟	10/15
职业素养		团队分工明确合理、操作规范、文明竞赛		5

四、竞赛方式

本竞赛为线下比赛，组队方式为团体赛，具体要求如下:

（一）参赛选手须为中等职业学校全日制在籍学生，五年制高职一至三年级（含三年级）学生也可报名参赛。凡在往届全国职业院校技能大赛中获一等奖的选手，不能再参加同一项目同一组别的比赛;

（二）每支参赛队由 3 名选手组成。参赛队可配指导教师，指导教师须为本校专兼职教师，每队限报 2 名指导教师。参赛选手和指导教师报名获得确认后不得随意更换;

（三）本赛项为单一场次，所有参赛队在现场根据给定的任务说明，在 6 小时内相互配合，采用小组合作的形式完成任务，最后以提交的结果文档作为最终评分依据。

五、竞赛流程

（一）竞赛时间表

表 2　竞赛时间表

日期	时间	内容
竞赛前两日	18:00 之前	裁判报到
	19:00—20:00	裁判工作会议
竞赛前一日	12:00 之前	各参赛队报到
	10:00—11:00	工作人员（含监考）培训会
	15:30—16:00	赛前领队会
	16:00—16:30	参赛队熟悉比赛场地
	17:00—18:00	现场裁判赛前检查，封闭赛场
竞赛当日	07:00—08:00	参赛队集合前往比赛现场
	08:00—08:10	赛场检录
	08:10—08:30	一次加密：参赛队抽取参赛编号
	08:30—08:45	二次加密：参赛队抽取赛位号
	08:45—09:00	参赛队进入比赛赛位，检查软硬件、题目发放
	09:00—15:00	竞赛进行
	15:00—15:20	收取各参赛队赛题及比赛结果文档
	15:00—17:00	申诉受理
	17:00—17:30	三次加密：竞赛结果等文件加密
	17:30—21:00	成绩评定与复核
	21:00—21:30	加密信息解密
	21:30—22:00	成绩汇总，报送及公布
竞赛后一日	09:00—10:00	闭赛式

（二）竞赛流程图

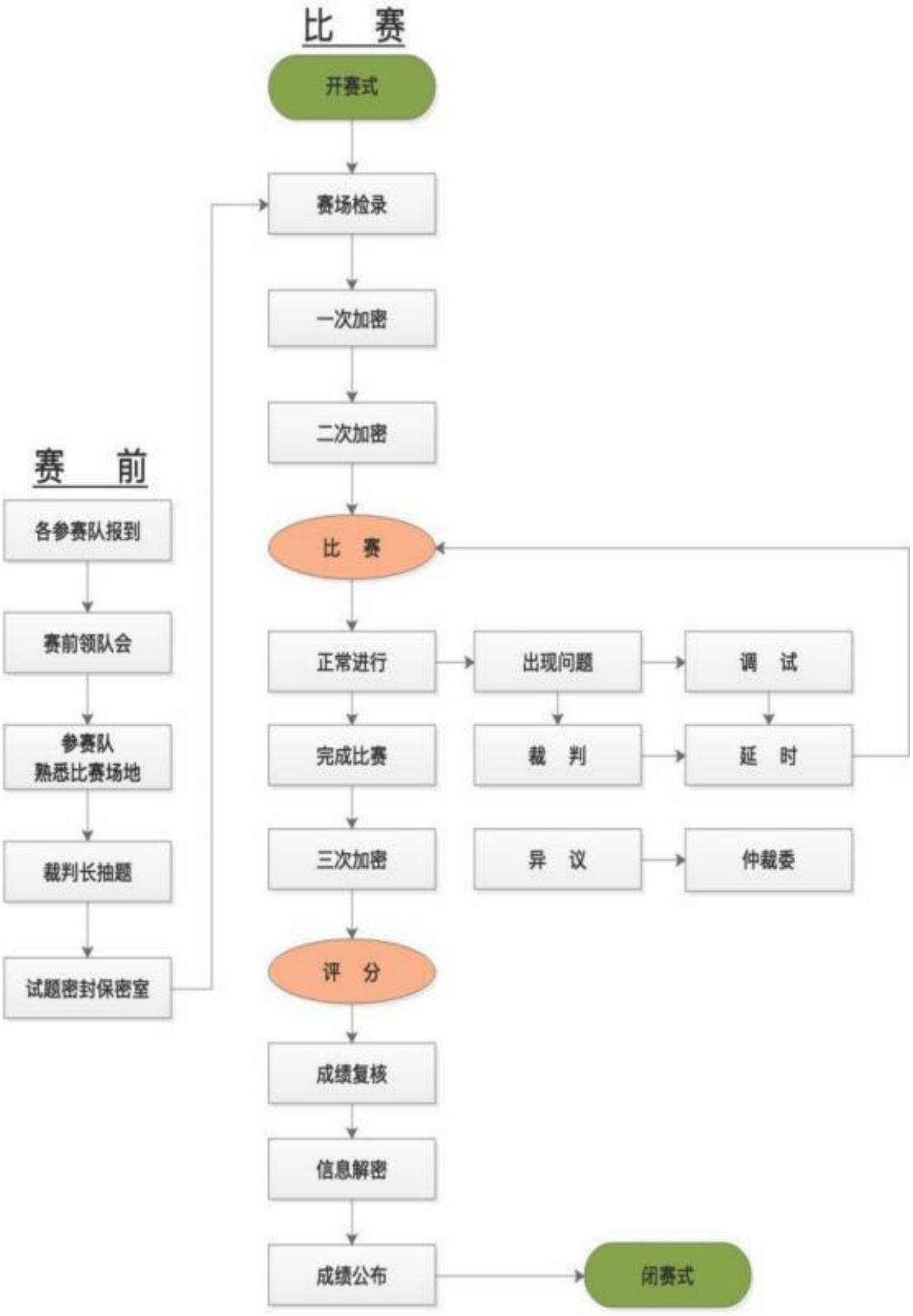

图 1 竞赛流程

六、竞赛规则

（一）选手报名:参赛选手须为中等职业学校全日制在籍学生，五年制高职一至三年级（含三年级）学生也可报名参赛。凡在往届全国职业院校技能大赛中获一等奖的选手,不能再参加同一项目同一组别的比赛;

（二）熟悉场地：竞赛前 1 日安排各参赛队领队、指导教师、参赛选手熟悉赛场;

（三）入场规则：参赛选手按规定时间到达指定地点，必须携带参赛证件，进行检录、一次加密、二次加密等流程，最终确定工位，选手迟到 10 分钟取消比赛资格。严禁参赛选手、赛项裁判、工作人员私自携带通讯、摄录设备进入比赛场地。参赛选手所需的硬件、软件和辅助工具统一提供,参赛队不得使用自带的任何有存储功能的设备，如手机、U 盘、移动硬盘等。参赛队在赛前领取比赛任务并进入比赛工位，比赛正式开始后方可进行相关操作;

（四）赛场规则：在比赛过程中，参赛选手如有疑问，应举手示意，现场裁判应按要求及时予以答疑。如遇设备或软件等故障，参赛选手应举手示意，现场裁判、技术人员等应及时予以解决。确因计算机软件或硬件故障，致使操作无法继续，经裁判长确认，予以启用备用设备。参赛选手不得因各种原因提前结束比赛。如确因不可抗因素需要离开赛场的，须向现场裁判员举手示意，经裁判员许可并完成记录后，方可离开。凡在竞赛期间内提前离开的选手，不得返回赛场;

（五）离场规则：比赛时间结束，选手应全体起立，结束操作。

参赛选手要确认已成功提交竞赛要求的文档，裁判员与参赛选手一起签字确认，经工作人员查收清点所有文档后方可离开赛场，离开赛场时不得带走任何资料；

（六）成绩评定

1.为尽量利用信息技术手段规避人为评分有主观性、公平性等因素，本赛项成绩评定建议采用机器自动评分、过程评分、结果评分三种组合评分方式。

2.竞赛采取三次加密。第一次加密裁判组织参赛选手第一次抽签，抽取参赛编号，替代选手参赛证等个人信息；第二次加密裁判组织参赛选手进行第二次抽签，确定赛位号，替换选手参赛编号；第三次加密裁判对各参赛队的赛项总结报告进行加密，替换赛位号。每个环节结束后，数据立即封存，加密裁判直接隔离，在评分结束后进行解密并统计成绩。

3.解密：裁判长正式提交评分结果并复核无误后，加密裁判在监督人员监督下进行三层解密：竞赛结果编号到工位号解密；工位号到参赛编号解密；参赛编号到参赛队名称解密。

4.抽检复核：为保障成绩评判的准确性，监督组对赛项总成绩排名前30%的参赛队伍的成绩进行复核；对其余成绩进行抽检复核，抽检覆盖率不得低于15%。监督组需将复检中发现的错误以书面方式及时告知裁判长，由裁判长更正成绩并签字确认。复核、抽检错误率超过 5%的，则认定为非小概率事件，裁判组需对所有成绩进行复核。

（七）其它未尽事宜，将在赛前向各领队做详细说明。

七、技术规范

本赛项以专业技术标准、行业技能标准、软件开发标准为准则。

（一）技术规范

表 3　专业技术标准

编号	标准号	标准名称
1	GB/T 11457-2006	信息技术、软件工程术语
2	GB8566-88	计算机软件开发规范
3	GB/T 12991-2008	信息技术数据库语言 SQL 第 1 部分: 框架
4	GB/T 21025-2007	XML 使用指南
5	GB/T 28821-1012	关系数据管理系统技术要求

表 4　大数据相关标准

编号	标准号	标准名称
1	GB/T 38672-2020	信息技术 大数据接口基本要求
2	GB/T 38673-2020	信息技术 大数据大数据系统基本要求
3	GB/T 38676-2020	信息技术 大数据存储与处理系统功能测试要求
4	GB/T 38643-2020	信息技术 大数据分析系统功能测试要求
5	GB/T 38675-2020	信息技术 大数据计算系统通用要求
6	GB/T 38633-2020	信息技术 大数据系统运维和管理功能要求
7	GB/T 38672-2020	信息技术 大数据接口基本要求
8	GB/T 38673-2020	信息技术 大数据大数据系统基本要求

表 5 软件开发标准

编号	标准号	标准名称
1	GB/T 8566-2001	信息技术软件生存周期过程
2	GB/T 15853-1995	软件支持环境
3	GB/T 14079-1993	软件维护指南
4	GB/T 17544-1998	信息技术软件包质量要求和测试

（二） 设备使用与操作规范

1. 计算机电源应保持良好，插座不得松动，发现有漏电现象应立即切断电源。

2. 开机前应检查有无异常情况。

3. 开机前先接电源、开外设，最后开主机。

4. 不能带电插拔外设及主机。

5. 如发现计算机有不正常现象时应立即停止操作，请裁判员检查后方可用机。

6. 做好数据资料的保密工作。

（三）操控人员应具备的专业知识

操控人员应具备数据采集与处理、数据分析与可视化、数据标注、大数据业务分析方法和方案架构、运行维护数据库系统等专业知识。

（四）操控人员应具备的技术技能

操控人员应具备分析系统数据来源、分析数据应用需求、设计数据资源整合解决方案、数据可视化、运行维护数据库系统、数据和信息处理等技术技能。

八、技术环境

（一）竞赛场地

赛场内设选手检录区、选手休息区、竞赛区、裁判组工作区、技术支持区、服务保障区、加密裁判封闭区 、配件仓库、保密室等，

1.选手检录区：选手等待检录、检录、一次加密、二次加密等职能区域;

2.选手休息区:选手检录前及竞赛离场休息、指导教师休息区域;

3.竞赛区：每个参赛队伍的比赛占地面积平均不少于 9 平方米。每个竞赛工位设工位编号，竞赛工位相对独立，确保选手独立开展竞赛，不受外界影响。赛场内安装摄像头，确保每个竞赛工位无盲区监控。

赛场环境的供电采用双强电设备，确保比赛用电的高可靠，各工位分区供电，强电弱电分开布线；场地采光、照明和通风良好，工位及竞赛桌面照度大于 500lux。赛项赛场整体平面布局如图 2 所示，其中竞赛区域详细赛位布局如图 3 所示。

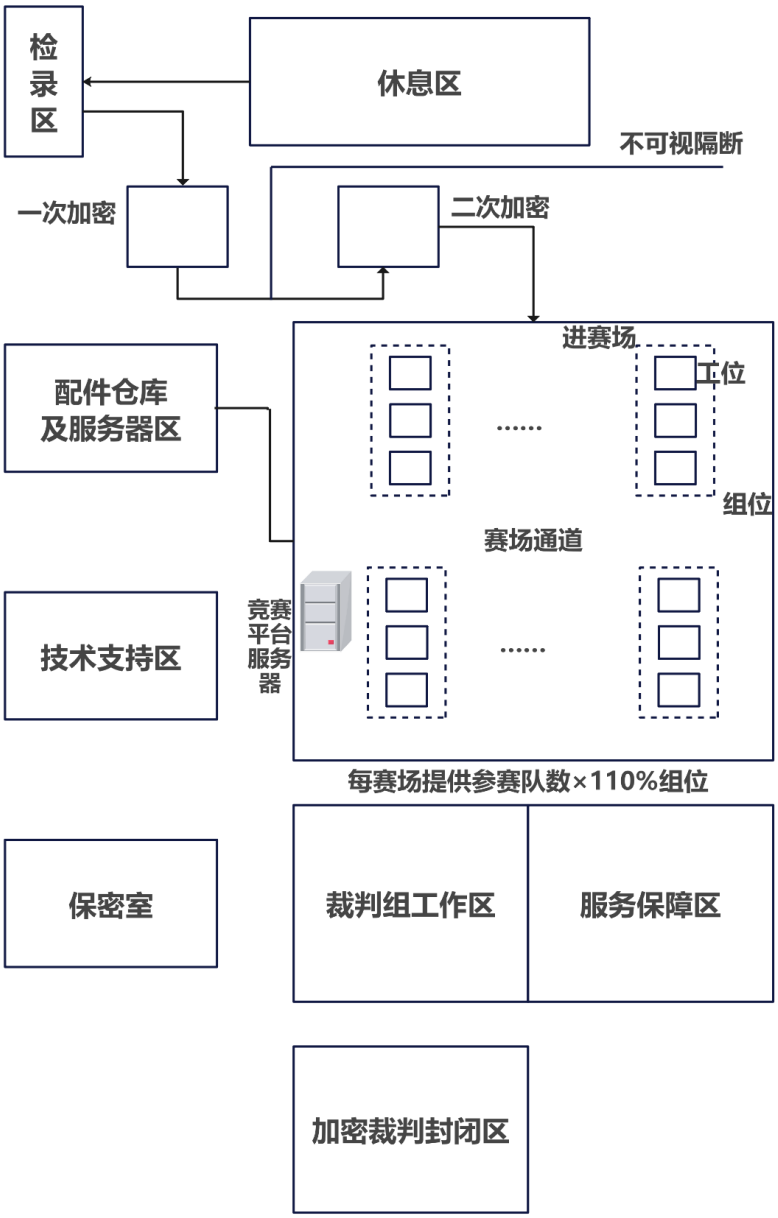

图2 赛项赛场平面整体布局图

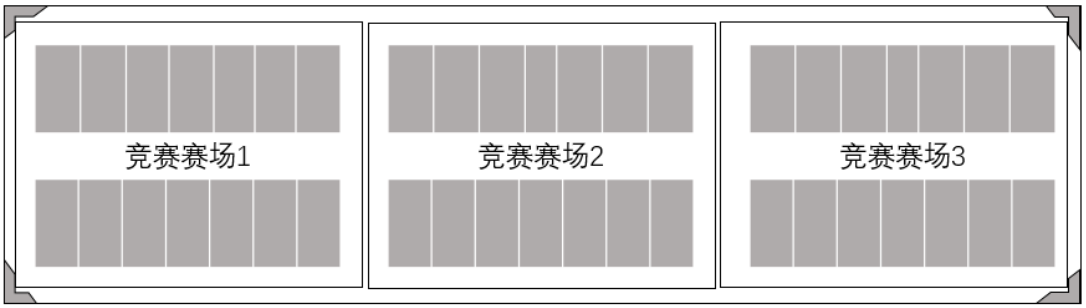

图3 赛项竞赛工位平面布局示意图

（1）整个比赛场地应保持通畅和开放，并配备防火防爆及其他安全设施。

（2）赛场周边设有卫生间、维修服务、医疗、生活补给站等公共服务区和紧急疏散通道，并在赛场周围设置隔离带。

（3）设立赛场开放区和安全通道，赛场走廊安装玻璃墙，透明、通亮，适合督巡，便于竞赛督查组巡视和竞赛接受采访，保证大赛安全有序进行和扩大社会影响力。

（4）场地配套提供稳定的水、电、气源和供电应急设备，并有保安、公安、消防、设备维修和电力抢险人员待命，以防突发事件。

（5）学生组、教师组技能竞赛在多工位进行。

（6）场内设施及布局

4.裁判组工作区：供裁判工作及休息，相对封闭，配讲台、办公桌、椅，电脑、投影仪、打印机等;

5.技术支持区：供技术支持人员工作及休息，为竞赛提供技术支持;

6.服务保障服务区:提供医疗等服务保障,以及竞赛期间备餐点,当地医疗机构要选派 2 名医务人员到赛场医疗点值班,处理比赛中突发情况问题;

7.加密裁判封闭区：一二次加密裁判在竞赛加密和解密期间，实现封闭管理所待区域，该区域不得提供任何与外界通讯途径;

8.配件仓库区：赛场所有实操易损配件（键盘、鼠标、网线等）、文具备货点，备件数量应达到赛场所需配件 10%，配件仓库严禁外人

进入，钥匙由裁判长掌握。

9. 保密室：保密室设在赛场附近，室内安装监控设备，安排专人值班，保密室钥匙由裁判长和监督组长分别掌握。

(二)技术平台

本赛项的硬件平台原则上采用通用技术实现。

表6 硬件平台

序号	设备名称	数量	说明
1	服务器	1	CPU：相当于或优于 2 颗 Intel Xeon Silver 4210 性能； 内存：不少于 128GB； 硬盘：不少于 1TB SSD； 网卡：至少双千兆网口；
2	PC 机	3	CPU：i5 及以上 内存：不少于 16GB 硬盘：不少于 500GB 显示器：1920*1080 及以上
3	交换机	1	8 口及以上千兆交换机

本赛项的软件平台原则上采用通用开源技术实现。

表7 软件平台

序号	软件类别	软件名称和版本或要求说明	单位	数量
1	机器评分系统	(1) 开源、免费； (2) 通用，与赛题无任何绑定； (3) 可检测虚拟机操作、程序运行、程序代码、并实现自动打分功能； (4) 可记录解题过程日志，且日志可溯源。	套	1
2	大数据集群操作系统	CentOS Linux release 7	套	1
3	大数据平台组件	Hadoop 2.7.7 以上	套	1
4		Yarn 3.1.3	套	1
5		ZooKeeper 3.4.14 以上	套	1
6		Hive 2.3.4/3.1.2	套	1

7		JDK 1.8	套	1
8		Scala 2.11 以上	套	1
9		Spark 3.0.0/3.1.1	套	1
10		Flume 1.9.0	套	1
11		Kafka 2.1 以上	套	1
12		Sqoop 1.4.7 以上	套	1
13		Flink 1.14.0	套	1
14		Redis 6.2.6	套	1
15		Excel 2016	套	1
16		HBase 2.2.3	套	1
17	关系型数据库	MySQL 5.7 及以上	套	1
18	开发语言	JavaScript	套	1
19		Java 1.8	套	1
20		html/html5+CSS	套	1
21		Python 3.6/3.7	套	1
22	开发库	numpy 1.18.5	套	1
23		pandas 1.3.4/2.1.3	套	1
24		matplotlib 3.5.0	套	1
25		Vue.js 3.2	套	1
26		ECharts 5.1 以上	套	1
27		pyecharts 2.0.4	套	1
28		snownlp 0.12.3	套	1
29		seaborn 0.11.2	套	1
30		openpyxl 3.0.9	套	1
31		lxml 4.9.3	套	1
32	开发工具	IDEA 2022（Community Edition）	套	1
33		PyCharm 2023(Community Edition)	套	1
34		HBuilderX 3	套	1
35		Visual studio code 1.79 以上	套	1
36	数据库工具	Navicat	套	1
37	浏览器	Chrome	套	1
38	SSH 连接工具	Xshell 或 MobaXterm 或系统自带终端工具	套	1
39	文档编辑器	WPS	套	1
40	数据采集	doccano 1.8.4	套	1
41	输入法	搜狗拼音输入法	套	1
42	PC 操作系统	Windows 10 64 位	套	1

九、竞赛样题

项目背景

近年来随着社会经济的快速发展，百姓生活水平的不断提高，外出旅游成为很多人生活的热门选择，如何促进旅游业的发展成为各级政府高度重视的工作。为了更好地统筹管理城市的旅游资源，某省的旅游管理部门采集了本省若干城市的酒店经营数据和用户评论数据，其中酒店经营数据包括日期、城市、酒店名称、酒店星级、酒店当天预定房间数、酒店当天入住客户数、酒店当天最高房价和酒店当天最低房价等字段，这些数据保存到文件 hotel.csv 中。用户评论数据包括日期、城市、酒店名称、住客评分、评论内容等字段，这些数据保存到文件 comments.csv 中。

你作为技术人员，需要通过数据采集清洗、数据标注、数据分析、数据可视化、业务分析等步骤对酒店经营数据和用户评论数据进行处理，从而为政府制定旅游发展的政策提供决策依据。请按照下面的要求完成相关任务。

模块一：平台搭建与运维

任务一：大数据平台搭建

子任务 1 Hadoop 完全分布式安装配置

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）从 Master 中的/opt/software 目录下将文件 hadoop-3.1.3.tar.gz 、 jdk-8u191-linux-x64.tar.gz 安装包解压到

/opt/module 路径中(若路径不存在，则需新建)，将命令和结果复制粘贴至对应报告中;

2）修改 Master 中/etc/profile 文件，设置 JDK 环境变量并使其生效，配置完毕后在 Master 节点分别执行“java -version”和“javac”命令，将命令和结果复制粘贴至对应报告中;

3)将三个节点分别命名为 master、slave1、slave2，并做免密登录，用 scp 命令并使用绝对路径从 Master 复制 JDK 解压后的安装文件到 slave1、slave2 节点(若路径不存在,则需新建),并配置 slave1、slave2 相关环境变量，将命令和结果复制粘贴至对应报告中;

4)在 Master 将 Hadoop 解压到/opt/module(若路径不存在,则需新建)目录下，并将解压包分发至 slave1、slave2 中，其中 master、slave1、slave2 节点均作为 datanode，配置好相关环境，初始化 Hadoop 环境 namenode，将命令和结果复制粘贴至对应报告中;

5)启动 Hadoop 集群（包括 hdfs 和 yarn），使用 jps 命令查看 Master 节点与 slave1 节点的 Java 进程，将命令和结果复制粘贴至对应报告中。

子任务 2 Hive 安装配置

本任务需要使用 root 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-hive-3.1.2-bin.tar.gz、mysql-connector-java-5.1.37.jar 安装包解压到/opt/module 目录下，将命令和结果复制粘贴至对应报告中。

2)设置 Hive 环境变量，并使环境变量生效，执行命令 hive --version 将命令和结果复制粘贴至对应报告中。

3)完成相关配置并添加所依赖包，将 MySQL 数据库作为 Hive 元数据库。初始化 Hive 元数据，并通过 schematool 相关命令执行初始化，将命令和结果复制粘贴至对应报告中。

任务二：数据库配置维护

子任务 1 在数据库中创建表

本任务在 MySQL 中创建表 t_comment 和表 t_hotel，并将用户评论数据 comments.csv 和酒店经营数据 hotel.csv 分别导入到表 t_comment 和表 t_hotel 中。具体要求如下:

1、创建用户评论表 t_comment，表 t_comment 的字段定义如下:

字段	类型	说明	备注
comment_date	date	日期	
city	varchar	城市	
hotel_name	varchar	酒店名称	
score	double	住客评分	
content	varchar	评论内容	

2、在 MySQL 中将 comments.csv 的数据导入表 t_comment。

3、创建酒店经营数据表 t_hotel,表 t_hotel 的字段定义如下:

字段	类型	说明	备注
current_date	date	日期	
city	varchar	城市	
hotel_name	varchar	酒店名称	
hotel_star	varchar	酒店星级	
rooms_booked	int	酒店当天预定房间数	
customers_checkedin	int	酒店当天入住客户数	
highest_price	int	酒店当天最高房价	
lowest_price	int	酒店当天最低房价	

4、在 MySQL 中将 hotel.csv 的数据导入表 t_hotel。

5、将以上 SQL 语句和运行结果复制粘贴至对应报告中。

子任务 2 使用 SQL 查询数据

本任务具体要求如下:

1、查询指定酒店的评论数量。

2、查询指定酒店的住客评分的平均值。

3、查询每个城市的酒店数量。

4、查询指定酒店的最高房价和最低房价。

5、将以上 SQL 语句和运行结果复制粘贴至对应报告中。

模块二: 数据获取与处理

任务一: 数据获取与清洗

子任务 1 对空字段数据进行处理

1、使用 python 读取 comments.csv 文件，将字段“酒店名称”为空的数据删除，并打印输出删除条目数，将打印内容粘贴至对应报告中，打印内容格式如下:

=== “删除酒店名称为空的数据共***条” ===

2、将字段“酒店名称”非空的数据保存到 comments1.csv 文件。

3、将符合题目要求的代码答案和 comments1.csv 的前 10 条记录数据复制粘贴至对应报告中。

子任务 2 对异常字段数据进行处理

住客评分的取值范围为[0,5]，其中 5 表示评价最高，0 表示评价最低。如果住客评分超出此取值范围的，都视为异常数据。本任务使用 pyhton 读取 hotel.csv 文件的数据，将字段“住客评分”异常

的数据删除,并打印输出删除条目数,将打印内容粘贴至对应报告中,打印内容格式如下:

=== “删除住客评分异常的数据共***条” ===

任务二：数据标注

本任务根据酒店的评论数据对酒店的类型打上标签,并将标签数据保存到指定位置。系统提前设定用户评价活跃阈值,如酒店的用户评价数量大于用户评价活跃阈值,则将该酒店的类型标注为“热门”,否则将该酒店的类型标注为“普通”，具体要求如下:

1、编写 python 程序读取读取 comments.csv 的数据，统计每个酒店的用户评价数量。

2、比较酒店的评价数量和用户评价活跃阈值，给该酒店的类型打上指定的标签（热门/普通），然后将打上标签的数据保存到 comments_tag.csv 中，comments_tag.csv 的字段定义如下:

酒店名称	评论数量	酒店类型
		热门/普通

任务三：数据统计

本任务使用 MapReduce 程序对酒店经营数据进行统计。

子任务 1 统计每个酒店的预订房间总数和入住客户总数

1）将 hotel.csv 文件上传至 HDFS 目录/hotel 中。

2）编译打包 MapReduce 程序，并将代码部署在 Hadoop 平台上运行，将程序运行结果保存到 HDFS 目录/result1 下。

3）读取 HDFS 目录/result1 的数据，将该数据复制粘贴至对应报告中。

子任务 2 统计每个城市不同星级酒店的数量

1）将 hotel.csv 文件上传至 HDFS 目录/hotel 中。

2）编译打包 MapReduce 程序，并将代码部署在 Hadoop 平台上运行，将程序运行结果保存到 HDFS 目录/result2 下。

3）读取 HDFS 目录/result2 的数据，将该数据复制粘贴至对应报告中。

模块三：业务分析与可视化

任务一：数据可视化

子任务 1 使用堆叠图展示城市星级酒店的数量

本任务使用堆叠图展示每个城市星级酒店的数量，本任务具体要求如下:

1）读取 hotel.csv，使用 pandas 分别统计每个城市的三星级酒店、四星级酒店和五星级酒店的数量。

2）使用 matplotlib 绘制堆叠图，堆叠图的标题为“各城市星级酒店的数量”,堆叠图的横坐标为城市名称,纵坐标为星级酒店数量。将可视化结果复制粘贴至对应报告中。

子任务 2 使用散点图展示各城市酒店入住客户总人数

将每个城市的所有酒店的入住客户的数量进行累加，就获得了每个城市入住客户的总人数。使用散点图展示不同城市入住客户的总人数,可以直观地对比这些城市的旅游接待能力,本任务具体要求如下:

1）读取 hotel.csv，使用 pandas 统计每个城市的所有酒店的入住客户总人数。

2）使用 matplotlib 绘制散点图，散点图的标题为“各城市酒店入住客户总人数”，将可视化结果复制粘贴至对应报告中。

子任务 3 使用柱状图展示酒店的评分数据

本任务使用柱状图展示酒店的评分数据，具体要求如下:

1）读取 hotel.csv，使用 pandas 统计分别统计三星级酒店、四星级酒店和五星级酒店的住客评分的平均值。

2）使用 matplotlib 绘制柱状图，柱状图的标题为“不同星级酒店的住客评分数据”， 柱状图的横坐标分别为三星级酒店、四星级酒店和五星级酒店，纵坐标为星级酒店对应的住客评分的平均值。柱状图为横向布局，将可视化结果复制粘贴至对应报告中。

任务二：业务分析

子任务 1 分析影响酒店入住客户数量的因素有哪些

结合模块三的任务一制作的可视化效果图，说明影响酒店入住客户数量的因素有哪些，并就如何提高酒店入住率给出相应的措施和建议。

子任务 2 分析影响酒店评分的因素有哪些

结合本模块三的任务一制作的可视化效果图，说明影响酒店评分的因素有哪些，并就如何提高酒店用户满意度和服务水平给出相应的措施和建议。

十、赛项安全

（一）比赛环境

1.赛场的布置，赛场内的器材、设备，应符合国家有关安全规定。

2.赛场周围要设立警戒线，防止无关人员进入，发生意外事件。

3.承办院校应提供保障应急预案实施的条件，明确制度和预案。

4.赛项执委会须会同承办院校制定开放赛场和体验区的人员疏导方案。

5.大赛期间，赛项承办院校须在赛场设置医疗医护工作站。

6.参赛选手、赛项裁判、工作人员严禁携带通讯、摄录设备和未经许可的记录用具进入比赛区域。

（二）生活环境

1.比赛期间，原则上由赛项承办院校统一安排参赛选手和指导教师食宿。承办院校须尊重少数民族参赛人员的宗教信仰及文化习俗。

2.比赛期间安排的住宿场所应具有旅游业经营许可资质。

3.大赛期间组织的参观和观摩活动的交通安全由赛区组委会负责。

（三）组队责任

1.各省、自治区、直辖市在组织参赛队时，须安排为参赛选手购买大赛期间的人身意外伤害保险。

2.各学校代表队组成后，须制定相关管理制度，并对所有选手、指导教师进行安全教育。

3.各参赛队伍须加强对参与比赛人员的安全管理，实现与赛场安

全管理的对接。

（四）应急处理

比赛期间发生意外事故，发现者应第一时间报告赛项执委会，同时采取措施避免事态扩大。赛项执委会应立即启动预案予以解决并报告赛区执委会。

十一、成绩评定

（一）评分队伍组成

成绩评定实行裁判长负责制，裁判组独立完成成绩评定工作。由竞赛裁判经验丰富的人员组成，大致组成如表 8，裁判和监督仲裁具体遴选要求参见《2023 全国职业院校技能大赛专家和裁判工作管理办法》中裁判遴选条件。

表 8 裁判员组成与执裁组成

序号	裁判员类别	人数
3	裁判长	1
4	检录裁判	2
5	加密解密裁判	3
6	现场评分裁判	5
8	评分裁判	10
合计		21

（二）评分标准制定原则

竞赛评分制定严格遵守公平、公正的原则，大数据应用与服务赛项评分采用赛项结果评分方法，始终贯彻落实竞赛一贯坚持的公平、公正和公开原则。

1.参与竞赛成绩管理的组织机构包括裁判组、监督仲裁组等，裁判组实行“裁判长负责制”，监督仲裁组组成结构、人数和相关要求参建《2023 全国职业院校技能大赛监督仲裁工作管理办法》中监督仲裁遴选条件和办法。

2.裁判评分方法，根据评分标准，各项目评分裁判根据选手操作过程和操作结果进行评分，独立评分。

3.成绩产生方法。为保证公开、公平、公正、透明地进行成绩评定，在裁判员的评分中，取两个评分裁判平均分作为选手技能得分。

4.成绩审核方法为各裁判员首先审核自身对选手的原始打分成绩，并签名，裁判长对所有裁判员的打分成绩进行审核，并签名，再由监督组对竞赛成绩抽检复核。

（三）评分方法

选手在完成任务之后，将任务完成结果拷贝至U盘中，由参赛选手队长签字确认（签工位号）。

评分采取分步得分、累计总分的计分方式。不计参赛选手的个人得分，只记录团体得分。

参赛队提交比赛任务结束请求或者在比赛时间终止后，不得再进行任何操作。否则，视为比赛作弊，给参赛队记警告一次。

在竞赛过程中，选手如有不服从裁判判决、扰乱赛场秩序、舞弊等不文明行为，由裁判长按照规定扣减相应分数并且给予警告，情节严重的取消竞赛资格，竞赛成绩记0分，队员退出比赛现场。

（四）成绩复核与解密

监督仲裁组将对赛项总成绩排名前 30%的所有参赛队伍（选手）的成绩进行复核;对其余成绩进行抽检复核,抽检覆盖率不得低于15%。如发现成绩错误以书面方式及时告知裁判长,由裁判长更正成绩并签字确认。复核、抽检错误率超过 5%的，裁判组将对所有成绩进行复

核。如有以上异常情况，应在专家组组长主持下，由裁判长带领裁判员、监督仲裁员共同处理。

成绩复核、确认无误后进行成绩排名，得出排名结果后进行解密，不允许先解密后排序。如无以上异常情况，成绩单由裁判长、监督仲裁长共同签字确认并封存直至公布成绩时开启。

（五）成绩公布方法

竞赛成绩经复核无误后，经裁判长、监督人员审核签字后，以纸质形式上报赛项组委会，并由赛项组委会最终公布结果为大赛最终公布成绩。成绩公示有效时间为 7:00-24:00，期间公示 2 小时后无异议，将赛项总成绩的最终结果录入赛务管理系统，经裁判长、监督仲裁组长在导出成绩单上审核签字后，在闭赛式上宣布；如公示结果有异议，由参赛队领队向赛项监督仲裁工作组递交亲笔签字同意的书面申诉报告。

（六）评分标准

表 9 评分标准

模块	任务	主要知识与技能点	分值
模块一：平台搭建与运维	任务一：大数据平台搭建	Hadoop 完全分布式下的 JDK 的解压安装、JDK 环境变量配置、节点配置、Hadoop 配置文件修改、运行测试等。	20
	任务二：数据库配置维护	使用 SQL 语句建立数据库和表。 使用 SQL 语句对表进行增删改操作。 使用 SQL 语句对表进行统计查询操作。 将 CSV 文件导入到数据库中。	10
	小计		30
模块二：数据获取与处理	任务一：数据获取与清洗	使用 pthon 程序读取 CSV 文件。 使用 pthon 程序处理空字段数据、异常字段数据。	10

	任务二：数据标注	使用 Python 对指定数据进行分类标注。 使用 Python 将标注后的数据保存到指定位置。	10
	任务三：数据统计	HDFS 上传 CVS 文件到指定目录下。 编写 MapReduce 程序对数据求解最大值、最小值、平均值和分区、分组操作等。 将计算结果保存到 HDFS 指定目录下。	15
	小计		35
模块三：业务数据分析与可视化	任务一：数据可视化	使用堆叠图展示相关统计数据 使用散点图展示相关统计数据 使用柱状图展示相关统计数据	20/15
	任务二：业务分析	理解业务场景，对数据报表进行分析研究，给出相应的建议和举措。	10/15
	小计		30
模块四：职业素养	考察职业素养	竞赛团队分工明确合理、操作规范、文明竞赛	5
	小计		5
总分			100

十二、奖项设置

本赛项奖项设团体奖。设奖比例为：以赛项实际参赛队总数为基数，一、二、三等奖获奖比例分别为 10%、20%、30%（小数点后四舍五入）。

如出现参赛队总分相同情况，依序按照模块二、模块一、模块三得分高低进行排名，在前序模块得分相同的情况下，按照后续模块得分排名。如果所有任务分值相同，则查看文档撰写规范、职业素养的分值进行排序。

获得一等奖的参赛队（团体赛）的指导教师授予“优秀指导教师奖”荣誉称号。

十三、赛项预案

赛场备用工位：赛场提供占总参赛队伍 10%的备用工位。

竞赛系统可靠性：竞赛系统使用的服务器应进行冗余，数据库、存储应使用高可用架构。提前开始运行，经过多次压力测试，由学校组织的真实竞赛环境测试。

竞赛备用服务器、客户机：现场提供占总参赛队伍 10%的备用服务器、客户机。

现场应急预案如下:

（1）服务器问题预案

若服务器在比赛过程中出现卡顿、死机等情况，参赛选手举手示意裁判，在裁判与技术支持人员确定情况后，可更换服务器。更换服务器的等待时间，可在比赛结束后延时。

（2）交换机问题预案

若交换机在比赛过程中出现传输速度慢或无故中断等情况，参赛选手举手示意裁判，在裁判与技术支持人员确定情况后，可更换交换机。更换交换机的等待时间，可在比赛结束后延时。

（3）PC 机问题预案

若 PC 机在比赛过程中出现死机、蓝屏等现象(重启后无法解决)，参赛选手举手示意裁判，在裁判与技术支持人员确定情况后，可更换备用工位或更换 PC 机进行答题。

十四、竞赛须知

（一）参赛队须知

1.参赛队名称统一使用规定的代表队名称。

2.参赛队员在报名获得审核确认后，原则上不再更换，如筹备过程中，选手因故不能参赛，所在学校需出具书面说明并按相关规定补充人员并接受审核；开赛前 10 日以内，参赛队不得更换参赛队员，允许缺员比赛。

3.参赛队按照大赛赛程安排凭大赛组委会颁发的参赛证和有效身份证件参加比赛及相关活动。

4.参赛队统一安排参加比赛前熟悉场地环境的活动。

5.各参赛队准时参加赛前领队会，领队会上举行抽签仪式抽取场次号。

6.各参赛队要注意饮食卫生，防止食物中毒。

7.各参赛队要发扬良好道德风尚，听从指挥，服从裁判，不弄虚作假。

（二）指导老师须知

1.各指导老师要发扬良好道德风尚，听从指挥，服从裁判，不弄虚作假。指导老师经报名、审核后确定，一经确定不得更换。

2.对申诉的仲裁结果，领队和指导老师应带头服从和执行，还应说服选手服从和执行。

3.指导老师应认真研究和掌握本赛项比赛的技术规则和赛场要求，指导选手做好赛前的一切准备工作。

4.领队和指导老师应在赛后做好技术总结和工作总结。

（三）参赛选手须知

1.参赛选手应遵守比赛规则，尊重裁判和赛场工作人员，自觉遵守赛场秩序，服从裁判的管理。

2.参赛选手应佩戴参赛证，带齐身份证、注册的学生证。在赛场的着装，应符合职业要求。在赛场的表现，应体现自己良好的职业习惯和职业素养。

3.进入赛场前须将手机等通讯工具交赛场相关人员保管，不能带入赛场。未经检验的工具、电子储存器件和其他不允许带入赛场物品，一律不能进入赛场。

4.比赛过程中不准互相交谈，不得大声喧哗；不得有影响其他选手比赛的行为，不准有旁窥、夹带等作弊行为。

5.参赛选手在比赛的过程中，应遵守安全操作规程，文明的操作。通电调试设备时，应经现场裁判许可，在技术人员监护下进行。

6.比赛过程中需要去洗手间，应报告现场裁判，由裁判或赛场工作人员陪同离开赛场。

7.完成比赛任务后，需要在比赛结束前离开赛场，需向现场裁判示意，在赛场记录上填写离场时间并签工位号确认后，方可离开赛场到指定区域等候评分，离开赛场后不可再次进入。未完成比赛任务，因病或其他原因需要终止比赛离开赛场，需经裁判长同意，在赛场记录表的相应栏目填写离场原因、离场时间并签工位号确认后，方可离开；离开后，不能再次进入赛场。

8. 裁判长发出停止比赛的指令，选手（包括需要补时的选手）应立即停止操作进入通道，在现场裁判的指挥下离开赛场到达指定的区域等候评分。需要补时的选手在离场后，由现场裁判召唤进场补时或比赛结束后自然延时补时。

9. 赛场工作人员叫到工位号、在等待评分的选手，应迅速进入赛场，与评分裁判一道完成比赛成绩评定。在评分过程中，选手应配合评分裁判，按要求进行设备的操作；可与裁判沟通，解释设备运行中的问题；不可与裁判争辩、争分，影响评分。

10. 遇突发事件，立即报告裁判和赛场工作人员，按赛场裁判和工作人员的指令行动。

（四）工作人员须知

1. 工作人员必须服从赛项组委会统一指挥，佩戴工作人员标识，认真履行职责，做好服务赛场、服务选手的工作。

2. 工作人员按照分工准时上岗，不得擅自离岗，应认真履行各自的工作职责，保证竞赛工作的顺利进行。

3. 工作人员应在规定的区域内工作，未经许可，不得擅自进入竞赛场地。如需进场，需经过裁判长同意，核准证件，有裁判跟随入场。

4. 如遇突发事件，须及时向裁判长报告，同时做好疏导工作，避免重大事故发生，确保竞赛圆满成功。

5. 竞赛期间，工作人员不得干涉及个人工作职责之外的事宜，不得利用工作之便，弄虚作假、徇私舞弊。如有上述现象或因工作不负责任的情况，造成竞赛程序无法继续进行，由赛项组委会视情节轻重，

给予通报批评或停止工作，并通知其所在单位做出相应处理。

（五）裁判员须知

1.裁判员执裁前应参加培训，了解比赛任务及其要求、考核的知识与技能，认真学习评分标准，理解评分表各评价内容和标准。不参加培训的裁判员，取消执裁资格。

2.裁判员执裁期间，统一佩戴裁判员标识，举止文明礼貌，接受参赛人员的监督。

3.遵守执裁纪律，履行裁判职责，执行竞赛规则，信守裁判承诺书的各项承诺。服从赛项专家组和裁判长的领导。按照分工开展工作，始终坚守工作岗位，不得擅自离岗。

4.裁判员有维护赛场秩序、执行赛场纪律的责任，也有保证参赛选手安全的责任。时刻注意参赛选手操作安全的问题，制止违反安全操作的行为，防止安全事故的出现。

5.裁判员不得有任何影响参赛选手比赛的行为，不得向参赛选手暗示或解答与竞赛有关的问题，不得指导、帮助选手完成比赛任务。

6.公平公正的对待每一位参赛选手，不能有亲近与疏远、热情与冷淡差别。

7.赛场中选手出现的所有问题如：违反赛场纪律、违反安全操作规程、提前离开赛场等，都应在赛场记录表上记录，并要求学生签工位号确认。

8.严格执行竞赛项目评分标准，做到公平、公正、真实、准确，杜绝随意打分；对评分表的理解和宽严尺度把握有分歧时，请示裁判

长解决。严禁利用工作之便，弄虚作假、徇私舞弊。

9.竞赛期间，因裁判人员工作不负责任，造成竞赛程序无法继续进行或评判结果不真实的情况，由赛项组委会视情节轻重，给予通报批评或停止裁判资格，并通知其所在单位做出相应处理。

十五、申诉与仲裁

（一）各参赛队对不符合赛项规程规定的设备、工具、材料、计算机软硬件、竞赛执裁、赛场管理及工作人员的不规范行为等，可向赛项仲裁组提出申诉，申诉主体为参赛队领队。

（二）仲裁人员的姓名、联系方式、工作地点应该在竞赛期间向参赛队和工作人员公示，确保信息畅通并同时接受大众监督。

（三）申诉启动时，由参赛队领队向赛项仲裁工作组递交亲笔签字同意的书面申诉报告。申诉报告应对申诉事件的现象、发生时间、涉及人员、申诉依据等进行充分、实事求是的叙述。非书面申诉不予受理。

（四）提出申诉应在赛项比赛结束后 2 小时内提出。超过 2 小时不予受理。

（五）赛项仲裁组在接到申诉报告后的 2 小时内组织复议，并及时将复议结果以书面形式告知申诉方。申诉方对复议结果仍有异议，可由领队向大赛仲裁工作组提出申诉。大赛仲裁工作组的仲裁结果为最终结果。

（六）申诉方不得以任何理由拒绝接收仲裁结果；不得以任何理由采取过激行为扰乱赛场秩序。仲裁结果由申诉人签收，不能代收；如在约定时间和地点申诉人离开，视为自行放弃申诉。

（七）申诉方可随时提出放弃申诉。

十六、竞赛观摩

为确保竞赛有序进行，以及公平公正，本赛项不设竞赛现场观摩区。

参赛队非竞赛人员及来访人员可通过休息室大屏幕的直播观摩比赛现场全过程。

十七、竞赛直播

（一）赛场内部署无盲点录像设备，能实时录制并播送赛场情况，本赛项竞赛时采用全过程录像；

（二）赛场外指导教师休息区有大屏幕或投影，同步显示赛场内竞赛状况；

（三）在不影响比赛的前提下，全过程、全方位安排现场直播，并设直播观摩区，让所有参赛教师和社会人员等观看比赛。赛后邀请媒体采访优秀选手、优秀指导教师、裁判专家或企业人士，突出赛项的技能重点与优势特色，为大赛宣传、资源转化提供全面的信息资料。视频资料也作为竞赛成果提交赛项执委会，作为竞赛历史材料供后续赛项提高进行参考，竞赛过程可作为教学资料进行资源转换，促进相关专业教学发展。

十八、赛项成果

大数据应用与服务资源转化工作由赛项执委会负责，依照《全国职业院校技能大赛赛项资源转化工作办法》的有关要求，通过多手段、全方位对赛项资源优秀成果进行转换，赛后向大赛执委会办公室提交大赛成果资源转化方案如下表，三个月内完成资源转化工作。

表 10 赛项成果推广表

一级资源项	二级资源项	内容简述	转化方式	进度安排	备注
风采展示	赛项宣传片	介绍大赛主题、目的、意义以及实施过程，突出展现参赛选手同台竞技的风采。	15 分钟视频	赛后 1 个月	1.承办校比赛当天全程摄像，拍摄比赛各个阶段 2. 承办校拍摄教师休息区和大屏展示区采集素材
	获奖选手风采展示片	介绍选手日常学习、备赛、参赛、获奖等环节的感受。	10 分钟视频	赛后 2 个月	1. 承办校闭幕式之后，访谈拍摄大赛一等奖参赛队 2.赛项执委会通知各个参赛队留存日常训练视频，比赛当天提供给承办单位用于剪辑
技能推广	训练大纲	介绍大赛训练过程要点	形成集训方案	赛后 3 个月	赛项专家组完成