**上海市“星光计划”**

**第十届职业院校职业技能大赛**

“大数据技术与应用”项目

**赛项规程**

上海市星光计划组委会竞赛办公室

2022年12月

目录

[1.项目简介 1](#_Toc1240331078)

[1.1项目描述 1](#_Toc620666517)

[1.2竞赛目的 1](#_Toc1214077740)

[1.3 相关文件 2](#_Toc1762446033)

[2.选手应具备的能力 2](#_Toc1188533560)

[3.竞赛模块及命题方式 5](#_Toc1938142173)

[3.1 竞赛模块 5](#_Toc1323543915)

[3.2 模块简述 5](#_Toc1166963779)

[3.2.1 模块A：大数据平台及组件搭建 5](#_Toc192085602)

[3.2.2 模块B：大数据平台管理与运维 6](#_Toc714791373)

[3.2.3 模块C：数据采集与处理 6](#_Toc475084693)

[3.2.4 模块D：大数据分析与挖掘 6](#_Toc404235705)

[3.2.4 模块E：数据可视化 6](#_Toc1498718474)

[3.3命题方式 7](#_Toc1125696855)

[3.4 命题方案 7](#_Toc256111915)

[4.评分规则 7](#_Toc915726817)

[4.1 测量分（客观） 7](#_Toc1752798917)

[4.3评分流程说明 8](#_Toc110728473)

[4.4 成绩公布方法 8](#_Toc1292607409)

[5.项目特别规定 9](#_Toc908150011)

[6.竞赛相关设施设备 9](#_Toc1110955648)

[6.1场地设备工具： 9](#_Toc1608748918)

[6.2材料： 10](#_Toc1423949096)

[6.3决赛选手须自备的设备和工具： 10](#_Toc754694304)

[6.4决赛场地禁止自带使用的设备和材料： 10](#_Toc1108748146)

[7.健康和安全 10](#_Toc1014484803)

[7.1赛场条件 10](#_Toc1573410488)

[7.2赛项保障 11](#_Toc196442658)

[7.3赛场布置 11](#_Toc929387567)

[7.4 参赛院校选手和教师的管理。 11](#_Toc1568273938)

[8.开放赛场 11](#_Toc1913276335)

[8.1公众要求 11](#_Toc15232167)

[8.2赛事宣传要求 11](#_Toc456476776)

[9.绿色环保 11](#_Toc1193587148)

本项目技术描述是对本竞赛项目内容的框架性描述，正式比赛内容及要求以竞赛当日公布的赛题为准。

# 1.项目简介

## 1.1项目描述

赛项竞赛内容的设计是适应国家大数据人才战略与行业发展需要，聚焦大数据技术与应用专业人才培养特点，覆盖大数据技术领域岗位的主要技能，围绕大数据技术领域的先进技术，主流产品及开发框架、标准化工作流程及工作步骤，力求突出面向应用的知识技能考核。

本赛项为团体赛，由3名选手组成。

赛项题目设计涵盖大数据平台及组件搭建、大数据平台管理与运维、数据采集与处理、大数据分析与挖掘、数据可视化等大数据主要技术领域，是对参赛选手大数据知识、技能和应用能力的综合检验，是从实际企业项目出发，完成大数据项目各阶段需要掌握的技能的综合考察，可促进高职院校大数据相关专业及课程建设，促进学生综合应用能力和职业素质的同步提高，提升毕业生的就业竞争力。

本赛项为电子与信息类赛项。

## 1.2竞赛目的

大数据技术与应用赛项是为适应大数据产业对高素质技术技能型人才的职业需求，覆盖大数据行业典型工作流程岗位，通过赛项丰富完善学习领域课程建设，使人才培养目标更贴近岗位实际需求，创新人才培养模式，建立健全多层次、多类型的大数据人才培养体系，提升专业建设服务社会和行业发展的能力。并借鉴世界技能大赛的先进理念，深化教学改革、推进校企合作，促进专业发展，展示师生风采。提高学生素养，培育工匠精神。发挥大赛社会效应，展示职业教育成果，服务上海产业转型发展，增强职业教育影响力和吸引力。

## 1.3 相关文件

本项目技术工作文件只包含项目技术工作的相关信息。除阅读本文件外，开展本技能项目竞赛还需配合其他相关文件一同使用：

1. **大数据竞赛智能评分系统使用文档**

大数据竞赛智能评分系统，是北京红亚科技开发的一套开放的大数据竞赛的评分系统，系统可免费提供使用，选手需要提前阅读竞赛智能评分系统的使用文档。

1. **大数据竞赛样题文件**

选手竞赛前需提前阅读竞赛样题，并完成竞赛样题测试，测试时间将统一时间开放；

1. **电脑使用文档**

选手掌握，并熟练使用PC电脑及windows10操作系统。

# 2.选手应具备的能力

|  |  |
| --- | --- |
| **模块** | **能力描述** |
| **A** | **大数据平台及组件搭建** |
|  | 个人需要知道和理解：   * 操作系统基本知识与概念 * 数据库的基本原理与逻辑关系 * 分布式网络环境的概念及校验过程 * Zookeeper的运行机制 * Hadoop集群分布式管理方式与规则 * Hbase列式数据库的结构 * Hive数据仓库的存储方式 * Spark计算引擎的使用方式及数据调用过程 |
|  | 个人应能够：   * 基础环境配置 * 数据库环境配置 * 分布式网络环境配置 * Zookeeper集群环境搭建 * Hadoop分布式集群搭建 * Hbase列式数据库搭建 * Hive环境配置 * Spark环境部署 |
| **B** | **大数据平台管理与运维** |
|  | 个人需要知道和理解：   * 大数据集群动态存储方式 * 大数据集群数据校验机制 * 大数据集群数据同步机制 * HDFS配置参数的意义及修改方式 |
|  | 个人应能够：   * 大数据集群配置管理 * 大数据集群数据节点管理 * 大数据集群监控服务配置与管理 |
| **C** | **数据采集与处理** |
|  | 个人需要知道和理解：   * 网络爬虫常见库使用 * 网站解析逻辑 * 掌握网页基本结构 * 利用chrome查看网页源码 |
|  | 个人应能够：   * 独立完成编写程序，构建爬虫需求 * 根据指定信息要求进行爬取数据项 * 将数据按照规定格式进行存储 * 将复杂的业务逻辑转化成合理的技术实现 * 良好的编码风格和编码习惯 |
| **D** | **大数据分析与挖掘** |
|  | 个人需要知道和理解：   * 典型数据聚类算法 * 典型数据降维算法 * 典型数据分类算法 * 线性回归算法 * 关联分析算法 * LDA机器学习算法 * K均值聚类机器学习算法 |
|  | 个人应能够：   * 编写典型数据聚类算法程序，并对数据进行分析与挖掘 * 编写典型数据降维算法程序，并对数据进行分析与挖掘 * 编写典型数据分类算法程序，并对数据进行分析与挖掘 * 编写线性回归算法程序，并对数据进行分析与挖掘 * 编写关联分析算法程序，并对数据进行分析与挖掘 * 编写LDA机器学习算法程序，并对数据进行分析与挖掘 * 编写K均值聚类机器学习算法程序，并对数据进行分析与挖掘 |
| E | **数据可视化** |
|  | 个人需要知道和理解：   * 典型数据可视化平台的使用方式 * 常见图表的绘制方式及参数意义 * 典型图表的应用场景及代表意义 * 图表间协同展现的逻辑关系 |
|  | 个人应能够：   * 使用标准化语句读取数据库或数据仓库中的数据 * 使用Echarts数据可视化图表库绘制折线图、饼形图、柱形图、直方图等图表 * 根据JSON语句规则，对数据进行交换格式转换 |

# 3.竞赛模块及命题方式

## 3.1 竞赛模块

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模块**  **编号** | **模块名称** | **竞赛时间**  **min** | **分数** | | |
| **评价分** | **测量分** | **合计** |
| A | 大数据平台及组件搭建 | 80 | 0 | 30 | 30 |
| B | 大数据平台管理与运维 | 30 | 0 | 15 | 15 |
| C | 数据采集与处理 | 30 | 0 | 15 | 15 |
| D | 大数据分析与挖掘 | 60 | 0 | 25 | 25 |
| E | 数据可视化 | 40 | 0 | 15 | 15 |
| 总计 |  | 240 | 0 | 100 | 100 |

竞赛整体时长为240分钟，各个模块的答题时间可根据作答情况自行调整，各个模块不单独交卷。如选手决赛成绩出现同分情况的，按照时间的排名顺序，早提交排名在前。

## 3.2 模块简述

## 3.2.1 模块A：大数据平台及组件搭建

本模块主要是考核学生对大数据平台及组件进行搭建部署的能力，主要包括对大数据平台基础环境配置、数据库环境配置、分布式网络环境配置、Zookeeper集群环境搭建、Hadoop分布式集群搭建、Hbase列式数据库搭建、Hive环境配置、Spark环境部署等组件的安装部署。

## 3.2.2 模块B：大数据平台管理与运维

本模块主要是考核学生对大数据平台管理与运维的能力，结合企业实际生产环境中的集群动态运维过程，对大数据集群进行的数据节点进行动态增加与删除，对集群的监控与管理，并根据现有数据存储的状态，对新增的数据储存节点进行数据冗余存储。

## 3.2.3 模块C：数据采集与处理

本模块主要是考核学生对数据的采集与处理能力，考察选手根据爬虫需求、对指定信息爬取数据项、将数据按照规定格式进行存储、将复杂的业务逻辑转化成合理的技术实现，并独立完成程序编写，根据数据存储要求，对数据进行有效存储，根据业务实际需要，对数据进行有效性、一致性、完整性处理。

## 3.2.4 模块D：大数据分析与挖掘

本模块主要是考核学生的大数据分析与挖掘能力，考察选手根据业务需求，结合典型数据聚类算法、数据降维算法、数据分类算法、线性回归算法、关联分析算法、机器学习算法等算法，实现算法代码程序的设计与编写，对数据进行分析与挖掘，实现数据价值的有效利用。

## 3.2.4 模块E：数据可视化

本模块主要是考核学生对数据的可视化展示能力，数据可视化是对数据最直观的展示，通过使用Echarts数据可视化图表库绘制折线图、饼形图、柱形图、直方图等图表，对数据进行多维度的展示，可有效展示出业务的现状、业务的关联关系、业务的运行规律等。

## 3.3命题方式

本项目竞赛题的命题方式：

本赛项为提前公布试题的项目，并于赛前 2 周公布样题，决赛真题包含样题 70%的内容。

## 3.4 命题方案

本项目竞赛命题是是从实际企业项目出发，将基础技能和工作任务融合的方式命题。赛项题目设计涵盖大数据平台及组件搭建、大数据平台管理与运维、数据采集与处理、大数据分析与挖掘、数据可视化等大数据主要技术领域，通过大数据主要技术在企业生产环境、企业实际项目中的应用，考察选手的综合应用能力。

竞赛设计思路在赛前2周内公布给参赛选手，针对典型技术点，由竞赛出题团队进行整体技术说明，在技术说明完成后，建立专用的交流群，对选手提出的各类问题进行解答，并对可能出现的变更予以及时公布。

# 4.评分规则

**裁判组构成**

1. 裁判长：由命题组组长担任，执行裁判长负责制。

2. 裁判员：各参赛学校可选派一名专业教师担任裁判员。

## 4.1 测量分（客观）

测量分（Measurement）打分方式：按模块设置若干个评分组，每组由2名及以上裁判构成。每个组所有裁判一起商议，在对该选手在该项中的实际得分达成一致后最终只给出一个分值。若裁判数量较多，也可以另定分组模式。

测量分评分准则样例表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类型 | 示例 | 最高分值 | 正确分值 | 不正确分值 |
| 满分或零分 | 按照虚拟机名称，修改对应主机名（分别为master、slave1、slave2，使用hostnamectl命令）  是否与答案匹配，代码及结果完全全匹配得分，不匹配不得分。 | 2 | 2 | 0 |

## 4.3评分流程说明

本竞赛项目模块A、模块B、模块C、模块D、模块E为过程评分，由专用的竞赛系统实时播报分数。如总分出现同分情况，以专用竞赛系统中前后排名为准，评分流程如下表：

|  |  |
| --- | --- |
| 模块 | 评分流程 |
| 模块A | 系统评分 分数复核 录入成绩 |
| 模块B | 系统评分 分数复核 录入成绩 |
| 模块C | 系统评分 分数复核 录入成绩 |
| 模块D | 系统评分 分数复核 录入成绩 |
| 模块E | 系统评分 分数复核 录入成绩 |

复核：由裁判组成员对分数进行审核，检查是否有误判，保证成绩真实有效；

录入：将裁判复核后的分数，录入至“评分表”内；

责任人：是监督和承担本竞赛项目全过程，为竞赛负总体责任；

## **4.4 成绩公布方法**

现场设立仲裁组，仲裁组由督考、裁判长和场地负责人组成。

裁判长和督考同时对成绩复核，并将参赛选手成绩汇总，各裁判员最终签字确认后，成绩经裁判长和督考确认后当场密封后报送大赛组委会，具体名次奖项由教委统一发文。

# 5.项目特别规定

1. 本竞赛项目，所有配套的文件及手册均采用简体中文编写。
2. 参赛选手应在竞赛前 15 分钟，凭竞赛抽签单和身份证进入考场。
3. 进入赛场后，参赛选手应按照抽签单进入指定工位，并检查设备状况，检查无误并 向裁判确认后方可开始竞赛。
4. 参赛选手应准时参赛，迟到 10 分钟以上，将不得入场，按自动弃权处理。
5. 参赛选手在竞赛期间可饮水、上洗手间，但其耗时一律计入竞赛时间。
6. 裁判发出开始竞赛时间信号后，参赛选手方可进行操作。
7. 参赛选手须独立完成所有项目，除征得裁判长许可，否则严禁与其他选手、与会人员、本单位裁判员交流接触。
8. 竞赛期间，参赛选手若遇到设备问题应向监考裁判举手示意，若为设备故障问题则停止计时，由工作人员进行处理。
9. 竞赛期间，选手只能访问自己竞赛账号，不得对尝试破解他人账号，不得向竞赛平台进行其它非常的访问或网络攻击，一经发现认定为考试作弊，取消本次选拔赛的参赛资格，考试成绩作废。
10. 参赛选手离开赛场前应保证所有电脑设备处于开机状态，如进行关机导致的一切后果，将由选手负责。
11. 监考裁判发出结束竞赛时间信号后，参赛选手应立即停止操作，依次有序离开赛场。

# 6.竞赛相关设施设备

## 6.1场地设备工具：

（*以每一个选手必须配备*）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 主体设备名称 | 型号 | 单位 | 数量 |
| 1 | 红亚大数据竞赛平台 | HonyaBigdata-match.AI | 套 | 1 |

## 6.2材料：

（*以每一个选手必须配备*）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 设备名称 | 型号 | 单位 | 数量 |
| 1 | PC电脑 | CPU：I5及以上；内存：4G及以上；硬盘：200G及以上；  配件：显示器、鼠标、键盘；  网络：具备千兆以上的局网环境； | 台 | 1 |

## 6.3决赛选手须自备的设备和工具：

本竞赛无需选手自备任何材料、设备和工具。

## 6.4决赛场地禁止自带使用的设备和材料：

|  |  |
| --- | --- |
| 序号 | 设备和材料名称 |
| 1 | 参赛选手禁止携带除抽签单、身份证、笔以外的任何物品进入竞赛场。 |

# 7.健康和安全

## 7.1赛场条件

（1）赛场布置，贯彻赛场集中，赛位独立的原则。选手竞赛单元相对独立，确保选手独立开展比赛，不受外界影响；赛位集中布置，保证竞赛氛围。

（2）卫生间、医疗、维修服务、生活补给站和垃圾分类回收点都在警戒线范围内，以确保大赛在相对安全的环境内进行。

（3）设置安全通道和警戒线，确保进入赛场的大赛参观、采访、视察的人员限定在安全区域内活动，以保证大赛安全有序进行。

## 7.2赛项保障

（1）建立完善的赛项保障组织管理机制，做到各竞赛单元均有专人负责指挥和协调，确保大赛有序进行。

（2）设置生活保障组，为竞赛选手与裁判提供相应的生活服务和后勤保障。

（3）设置技术保障组，为竞赛设备、软件与竞赛设施提供保养、维修等服务，保障设备的完好性和正常使用，保障设备配件与操作工具的及时供应。

（4）设置医疗保障服务站，提供可能发生的急救、伤口处理等应急服务。

（5）设置外围安保组，对赛场核心区域的外围进行警戒与引导服务。

## 7.3赛场布置

（1）赛场应进行周密设计，绘制满足赛事管理、引导、指示要求的平面图。竞赛举行期间，应在竞赛场所、人员密集的地方张贴。

（2）赛场平面图上应标明安全出口、消防通道、警戒区、紧急事件发生时的疏散通道。

（3）赛场的标注、标识应进行统一设计，按规定使用大赛的标注、标识。赛场各功能区域、赛位等应具有清晰的标注与标识。

## 7.4 参赛院校选手和教师的管理。

赛事承办院校统一安排参赛选手和指导教师午餐。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，根据国家相关的民族、宗教政策安排好少数民族参赛选手和教师的饮食。

# 8.开放赛场

## 8.1公众要求

（1）赛场内除指定的监考裁判、工作人员外，其他与会人员须经组委会同意或在组委会负责人陪同下，佩带相应的标志方可进入赛场。

（2）允许进入赛场的人员，只可在安全区内观摩竞赛。

（3）允许进入赛场的人员，应遵守赛场规则，不得与选手交谈，不得妨碍、干扰选手竞赛。

（4）允许进入赛场的人员，不得在场内吸烟。

## 8.2赛事宣传要求

（1）承办单位应极力邀请报刊媒体、网络媒体、电视媒体等媒体参与赛事宣传活动，提高赛事知名度。

（2）媒体记者必须经组委会同意并佩戴相应的标志方可进入赛场。

（3）媒体记者进入赛场后，应遵守赛场规则，不得与选手交谈，不得妨碍、干扰选手竞赛。

# 9.绿色环保

（1）赛场严格遵守我国环境保护法。

（2）赛场所有废弃物应有效分类并处理，尽可能地回收利用。