**2025年唐山市中等职业学校学生技能大赛**

**大数据应用与服务 赛项技术规程**

**一、赛项名称**

赛项名称：大数据应用与服务

赛项组别：中职组

归属专业大类：电子与信息大类

**二、竞赛目的**

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，从而有效促进大数据教学模式的探索性改良，推进相关专业课程体系、教学内容和教学方法等教学资源的质量提升和丰富完善，进而推动大数据专业教育上层建筑体系质的飞跃。

**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竞赛内容的时长、分值

|  |  |  |
| --- | --- | --- |
| 模块 | 主要内容 | 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。 | 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 | 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。 | 10 |
| 任务二： 数据标注 | 使用Python语言对数据进行分 类标注。 | 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。 | 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。 | 20 |
| 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。 | 10 |
| 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。

（二）每校参赛队伍不超过两队，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，每队指导教师不超过两人，每个学校设领队1人。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4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）竞赛操作流程

大数据应用与服务赛项的竞赛流程如图1所示。



**图1 赛事流程**

（二）竞赛日程安排

竞赛日程表（仅供参考）

|  |  |  |
| --- | --- | --- |
| **日期** | **时间** | **事项** |
| 竞赛日 | 07:30—08:00 | 赛场检录 |
| 08:00—08:10 | 一次加密：参赛队抽取参赛编号 |
| 08:10—08:20 | 二次加密：参赛队抽取赛位号 |
| 08:20—08:30 | 参赛队进入比赛赛位，进行赛前设备、材料检查 |
| 08:30—12:30 | 选手比赛 |
| 12:30—14:30 | 申诉受理 |
|  | 成绩核定和解密 |

**表1 竞赛时间一览表**

**六、竞赛规则**

（一）参赛资格：参赛选手应为全日制正式学籍的中职在校学生（含技工学校）或五年制高职一、二、三年级学生。

（二）报名要求：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，队员因故不能参赛，所在学校出具书面说明并按相关规定补充人员并接受审核；竞赛开始后，参赛队不得更换参赛队员，允许队员缺席比赛。

（三）检录：由检录工作人员依照检录表进行点名核对，并检查确定无误后向评委递交检录单。

（四）引导：参赛选手凭证件进入赛场，不得携带其他显示个人身份信息的物品，不得携带与竞赛无关的电子设备、通讯设备及其他相关资料与用品。现场评委负责引导参赛队伍至赛位前等待竞赛指令。比赛开始前，在没有评委允许的情况下，严禁随意触碰竞赛设施和阅读赛卷内容。比赛中途不得离开赛场。

（五）竞赛开始：由首席评委宣布比赛开始，各参赛队开始比赛。

（六）竞赛过程：竞赛过程中，如遇设备或软件等故障，参赛选手应及时举手示意。现场评委、技术人员等应及时予以解决。确因计算机软件或硬件故障，致使操作无法继续的，经现场工作人员同意，予以启用备用计算机。如遇身体不适，参赛选手应及时示意，现场医务人员按应急预案救治。如有其它问题，参赛选手应举手示意评委，评委应按照有关要求及时予以答疑。

（七）比赛结果：最终成绩单，经评委组各评委签字、大赛组委会审核后进行公布。公布时间为2小时。成绩公布无异议后，本赛项成绩正式生效。

（八）申诉与仲裁：本赛项在比赛过程中若出现有失公正或有关人员违规等现象，参赛队领队可在成绩公布后2小时之内向大赛组委会提出书面申诉。口头报告或其他人员要求解释处理，组委会不予受理。

书面申诉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，并由领队亲笔签名。非书面申诉不予受理。

大赛组委会在接到申诉报告后的2小时内组织评委团队复议，并及时将复议结果告知申诉方。申诉方可随时提出放弃申诉。

**七、竞赛环境**

竞赛场地包括竞赛机房、裁判机房及相关区域。

**八、技术平台**

技术平台

|  |  |  |
| --- | --- | --- |
| 序号 | 品名 | 规格要求说明 |
| 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。 |
| 2 | 大数据集群操作系统 | CentOSLinux release 7 |
|
| 3 | 大数据平台组件 | Hadoop 2.7.7以上 |
| Yarn 3.1.3 |
| ZooKeeper 3.4.14以上 |
| Hive 2.3.4/3.1.2 |
| JDK 1.8 |
|  Scala 2.11以上 |
|  Spark 3.0.0/3.1.1 |
|  Flume 1.9.0 |
|  Kafka 2.1以上 |
|  Sqoop 1.4.7以上 |
|  Flink 1.14.0 |
|  Redis 6.2.6 |
|  Excel 2016 |
|  HBase 2.2.3 |
| 4 | 关系型数据库 | MySQL 5.7及以上 |
| 5 | 开发语言 | JavaScript |
| Java 1.8 |
| html/html5+CSS |
| Python 3.6/3.7 |
| 6 | 开发库 | numpy 1.18.5 |
| pandas 1.3.4/2.1.3 |
| matplotlib 3.5.0 |
| Vue.js 3.2 |
| ECharts 5.1以上 |
| pyecharts 2.0.4 |
| snownlp 0.12.3 |
| seaborn 0.11.2 |
| openpyxl 3.0.9 |
| lxml 4.9.3 |
| 7 | 开发工具 | IDEA2022（Community Edition） |
| PyCharm2023(CommunityEdition) |
| HBuilderX 3 |
| Visualstudio code 1.79以上 |
| 8 | 数据库工具 | Navicat |
| 9 | 浏览器 | Chrome |
| 10 | Ssh连接工具 | Xshell或MobaXterm或系统自带终端工具 |
| 11 | 文档编辑器 | WPS或office |
| 12 | 数据采集 | doccano 1.8.4 |
| 13 | 输入法 | 搜狗拼音输入法 |
| 14 | PC操作系统 | Windows 10 64位 |

九、奖项设定

本赛项奖项为团体奖，团队成员授予相应奖项，团体奖按团体成绩由高到低排序，以参赛队伍数量为基数：一等奖占比10%，二等奖占比20%，三等奖占比30%。（小数点后四舍五入）