

2022年度“强国杯”技术技能大赛

——大数据技术应用赛项

一、大赛范围、赛题类型和其他

(一) 命题原则

此次竞赛设计完全与工作实需求为背景，按照大数据工程技术人员国家职业技能标准，聚焦大数据技术应用技术前沿趋势，围绕大数据环境部署与应用、数据采集与治理、大数据分析与挖掘、数据分析算法与数据可视化技术应用等方面进行重点考察，突出企业所需专业技能及新技术应用，遵循竞赛一贯坚持的公开、公平、公正原则，突出职业能力考核及工匠精神要求。

(二) 考核范围、题型、考核时间

阶段	时长	模块	模块名称及考核内容
资格赛	120分钟	A	理论知识模块 <ul style="list-style-type: none">• 分析工具知识• 数据库知识• 数据分析知识• 数据管理知识• 数据安全知识• 数据保密知识• 相关法律知识
		B	EXCEL/WPS 工具数据分析 <ul style="list-style-type: none">• 工具设置及快捷键使用• 数据分列

半决赛	180分钟		<ul style="list-style-type: none"> • 数据筛选 • 统计分析函数/文本处理函数 • 数值运算函数/逻辑判断函数 • 日期计算函数/匹配查找函数 • 使用函数实现报表开发 • 数据透视基础表制作
		C	SQL 数据分析
			<ul style="list-style-type: none"> • DDL/DML/DQL/DCL 使用 • SQL 查询关键词 • SQL 核心框架 • SQL 书写顺序 • 简单查询 • 条件查询 • 多表查询
		D	Python 数据分析
			<ul style="list-style-type: none"> • Python 基础语法使用 • Numpy 数据计算 • Pandas 数据分析 • Matplotlib 数据可视化
		E	伪分布式集群搭建
			<ul style="list-style-type: none"> • Linux 操作系统 • 主机名/防火墙设置 • 系统服务开启/关闭 • Unix 系统文本编辑器使用 • SSH 远程登录 • JDK 安装部署 • Hadoop 伪分布式部署
		F	理论知识模块
			<ul style="list-style-type: none"> • 大数据基本概念 • Hadoop 基础知识 • Hive 基础知识 • Spark 生态体系 • 数据迁移知识 • 数据采集知识 • 离线处理与实时处理 • 大数据平台运维基础知识
			G
	大数据平台运维		
<ul style="list-style-type: none"> • 集群配置与管理 • 大数据各组件常见参数配置 • 查看组件相关日志 • 精准故障判断 			

			<ul style="list-style-type: none"> • 标准化运维维修 • 集群本地测试 		
		H	数据迁移 <ul style="list-style-type: none"> • 数据传输工具 Sqoop 环境部署 • 传统数据库数据导入大数据集群 • 大数据集群数据出到传统数据库 		
		I	数据采集 <ul style="list-style-type: none"> • 使用爬虫代码采集数据 • 使用 Flume 进行日志采集 		
		J	离线数据处理 <ul style="list-style-type: none"> • 编写 MapReduce 进行数据分析 • 使用 Hive SQL 进行数据分析 		
		K	实时数据处理 <ul style="list-style-type: none"> • Kafka 集群部署 • 使用 Kafka 消费数据 • 使用 Spark 进行数据处理 • 使用 HBase 进行实时读写 		
		L	数据可视化 <ul style="list-style-type: none"> • 基于前端进行数据可视化展示 • 使用工具进行数据可视化展示 		
		M	Spark MLlib <ul style="list-style-type: none"> • 基本统计分析 • 特征抽取 • 特征变换 • 分类与回归 		
		全国总决赛	240 分钟	N	集群优化 <ul style="list-style-type: none"> • 操作系统调优 • HDFS 参数调优 • MR 参数调优 • YARN 参数调优 • 应用系统调优 • 集群数据均衡 • 集群读写性能测试
				O	开源任务调度系统应用 <ul style="list-style-type: none"> • 工作流调度实现方式 • 工作流定义 • 工作流传参 • 定时执行 • 资源管理

		<ul style="list-style-type: none"> • workflows执行 • workflows管理
	P	<p>运用 Python 编写算法对文本类数据案例分析</p> <ul style="list-style-type: none"> • 文本数据获取 • 分词/词向量 • 构造词频矩阵 • 逻辑回归/贝叶斯分类 • 分类结果可视化
	Q	<p>运用 Python 编写算法对图像类数据案例分析</p> <ul style="list-style-type: none"> • 图像采集 • 图像增强 • 灰度/二值化处理 • 边缘/轮廓检测 • 傅立叶变换/霍夫变换 • 目标识别

(三) 考核方式

竞赛各赛段要求以公开、公平、公正原则开展，采用系统智能评测，综合考虑时间因素，竞赛环节中实时公开成绩，公开选手动态，实时公开选手排行榜。考核方式采用先进的智能化测评技术，将所有竞赛结果进行智能评测，全面体现公开、公平、公正原则，通过云答题平台的创新的自动化监测技术，自动识别参赛选手的平台搭建是否成功、自动判断数据分析结果、自动评估算法模型识别精度等，全程采用自动化智能判断。

(四) 命题方式

本竞赛命题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。赛项题目的设计，聚焦大数据技术应用技术前沿趋势，围绕大数据环境部署与应用、数据采集与治理、大数据分析与挖掘、

数据分析算法与数据可视化技术应用等方面进行重点测评，考察选手的综合应用能力。