

2021 年甘肃省高职院校技能大赛（高职组） “大数据技术与应用”赛项竞赛规程

一、赛项名称

赛项名称：大数据技术与应用

赛项组别：高职组

竞赛形式：团体赛

赛项专业大类：电子信息

二、竞赛目的

为适应大数据产业对高素质技术技能型人才的职业需求，赛项以大数据技术与应用为核心内容和工作基础，重点考查参赛选手基于 Hadoop 平台环境下，充分利用 Hadoop 技术生态组件的特点，综合软件开发相关技术，解决实际问题的能力，激发学生对大数据相关知识和技术的学习兴趣，提升学生职业素养和职业技能，努力为中国大数据产业的发展储备及输送新鲜血液。

通过举办本赛项，可以搭建校企合作的平台，提升大数据技术与应用专业及其他相关专业毕业生能力素质，满足企业用人需求，促进校企合作协同育人，对接产业发展，实现行业资源、企业资源与教学资源的有机融合，使高职院校在专业建设、课程建设、人才培养方案和人才培养模式等方面，跟踪社会发展的最新需要，缩小人才培养与行业需求差距，引领职业院校专业建设与课程改革。

三、竞赛内容

赛项以大数据技术与应用为核心内容，重点考查参赛选手基于

Hadoop 平台环境下，利用 Hadoop 技术生态组件，综合软件开发相关技术，解决实际问题的能力，具体包括：

1. 掌握按照项目需求配置管理 Hadoop 大数据平台及相关生态组件；
2. 掌握网络爬虫的相关技术，完成指定数据的采集及处理能力；
3. 综合利用 Spark 技术、数据仓库 Hive 工具完成数据清洗、分析及数据推送等一系列大数据操作；
4. 综合运用 HTML、Python 等开发语言，对数据进行可视化呈现；
5. 根据数据可视化结果，完成数据分析报告的编写；
6. 根据竞赛任务要求，参赛队伍在 4 小时竞赛时间内须按顺序完成竞赛任务，但每项任务用时可自行掌握。

表 1：竞赛内容构成

考核环节	考核知识点和技能点	描述
Hadoop 平台及组件的部署管理	Hadoop 平台安装部署和基本配置	考察 Hadoop 平台及组件的部署能力，掌握常用的基本配置和命令，能够部署和管理 Hadoop 高可用集群。
	Hadoop 集群节点的动态增加与删除	
	Hadoop 平台相关组件部署与管理	
	Hadoop 平台的高可用	
数据采集	使用开发者工具查看网页源码，分析网页结构，明确数据采集对象	考察学生多维度数据采集能力，包括对关系型数据库、非关系型数据库和网络爬虫技术的应用。
	构建数据采集请求，抓取网络数据	
	利用网络爬虫相关组件实现网络数据爬取	
	规则文件数据和关系型数据库数据抓取以及数据同步	
	非关系型数据库数据抓取以及数据同步	
	数据采集结果导出及数据库推送	
数据清洗与分析	基于 Hadoop 平台架构组件和多维度的数据采集，实现数据一致性检查、无效值和缺省值的处理	考察对分布式计算、分布式存储系统、数据仓库等综合应用能力，使

	多表数据合并和离群值处理	用 Spark 技术、数据仓库 Hive 工具完成数据清洗、分析及数据推送等一系列大数据操作
	通过常见的数据分析算法，对数据进行标准化、离散化和多元化分析	
	掌握数据仓库导入、导出，利用数据仓库相关命令或代码实现数据多维度、多层次的分析	
	对数据的查询、整理和计算。进行编译、打包、发布，执行程序，完成数据处理、清洗。	
	实现不同数据库间的文件传输及转换	
	数据预测分析	
数据可视化	编写后台代码实现数据库访问和数据整理	通过常见的数据可视化方法，将数据分析结果以图表的形式进行呈现，使用 Python 编程语言，实现数据源分析结果展现
	编写 Web 前端代码，对数据分析结果进行呈现	
综合分析	通过知识技能，根据数据分析、预测及可视化结果进行分析，做出分析报告。	考察学生对大数据技术与分析的综合操作能力和业务分析能力

四、竞赛方式

最终以省上下发文件为准执行。

五、竞赛试题

本赛项样题详见《附件一》。

六、竞赛环境

竞赛现场设置场内竞赛区、现场裁判工作区、技术支持区等。

监考人员协助裁判长和现场裁判做好负责工位范围内的秩序维持，监考人员不得在考场内随意走动。

技术支持保障人员在技术支持服务区候场，有需要时在现场裁判

的带领下到相关的工位进行赛场技术支持保障，在条件具备时，技术支持区可不设置在考场内。

场内竞赛区为参赛队提供统一的竞赛设备；无需选手自带任何工具及附件。

竞赛现场符合消防安全规定，现场消防器材和消防栓合格有效，应急照明设施状态合格，赛场明显位置张贴紧急疏散图，赛场地面张贴荧光疏散指示箭头，赛场出入口专人负责，随时保证安全通道的畅通无阻。

各工位分区供电，强电弱电分开布线，现场临时用电满足《施工现场临时用电安全技术规范》JGJ46-2005 的要求。竞赛现场通风良好、照明需符合教室采光规范。

七、技术规范

本赛项的技术规范将包括：相关专业的教育教学要求、行业、职业技术标准，以及根据高职目录修订后的大数据技术与应用相关专业人才培养标准和规范，适时地修订本赛项遵循的技术规范。

表 2：基础标准

标 准	内 容
GB/T 11457-2006	信息技术、软件工程术语
GB8566-88	计算机软件开发规范
GB/T 12991-2008	信息技术数据库语言 SQL 第 1 部分：框架
GB/T 21025-2007	XML 使用指南
GB/T 20009-2005	信息安全技术数据库管理系统安全评估准则 已发布
GB/T 20273-2006	信息安全技术数据库管理系统安全技术要求
20100383-T-469	信息技术安全技术信息安全管理体系实施指南

表 3：软件开发标准

标 准	内 容
GB/T 8566 -2001	信息技术 软件生存周期过程

GB/T 15853 -1995	软件支持环境
GB/T 14079 -1993	软件维护指南
GB/T 17544-1998	信息技术 软件包 质量要求和测试

八、技术平台

表 4：竞赛设备

设备类别	数量	设备用途	基本配置
竞赛服务器	4 台。 采用集群管理方式； 1 台备用	支撑大数据竞赛管理系统运行使用。内嵌虚拟化资源管理控制端，作为虚拟化资源管理系统的计算资源、网络资源和存储资源的源节点。	1、CPU 模块：2*Intel 5118(2.3GHz/12 核/16.5MB/105W) 2、内存模块：8*32GB 2Rx4 DDR4-2666P-R 3、硬盘模块：6*600GB 12G SAS 10K 2.5in EP 512n 4、RAID 卡：1*12Gb 2 端口 SAS RAID 卡(带 2GB 缓存,支持 8 个 SAS 口,PCIe,不含超级电容) 5、网口：4 端口千兆电接口网卡-360T-B2 6、电源模块：550W 交流电源模块(白金) 7、超级电容模块：LSI G3 超级电容模块(适配 2U 机型)
大数据竞赛平台	1	支撑大数据竞赛实操考试	系统基于 kvm 构建，可模拟大数据环境搭建、大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘、大数据展现和应用等贯穿大数据技术的相关知识点，提供大数据竞赛管理系统所需的虚拟服务器，结构化、半结构化及非结构化数据的数据库等基础支撑环境；涵盖分布式虚拟存储技术，大数据获取、存储、组织、分析和决策操作的可视化技术。具体包括：Hadoop、HDFS、Hbase、Hive、

			MapReduce、Kafka、Spark、Storm、Mahout、MySQL、Echarts 等，所涉及开发语言包括 Java、Python、Scala、HTML、Javascript 等。
客户端	每支参赛队伍 3 台。 根据参赛团队数量，配备 10% 的备份机器。	竞赛选手比赛使用。	性能相当于 i5 处理器，8G 以上内存，SSD 200G 以上硬盘，显示器要求 1024*768 以上。

表 5：软件环境

设备类型	软件类别	软件名称、版本号
服务器集群	大数据集群操作系统	CentOS 7.4
	大数据分析平台组件	Hadoop 2.6.0
		Yarn 2.6.0
		Zookeeper 3.4.5
		Flume 1.6.0
		Sqoop 1.4
		kafka 1.0
		Spark 2.0
	数据库	MySQL 5.7
开发客户端	PC 操作系统	Windows 10 64 位
	浏览器	Chrome、Google
	开发语言	Python 3.6 64bit
		Scala 2.11
	开发工具	Pycharm 2019 (Community Edition)
		IDEA 2019 (Community Edition)
	数据采集组件	Scrapy
	数据可视化组件	ECharts
		Flask

		Jinja2
	文档编辑器	MS Office 2016
	输入法	拼音输入法

九、成绩评定

（一）评分方法

选手在完成比赛任务之后，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、错误不传递、累计总分的计分方式，只记录参赛队整体成绩。

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。

（二）评分标准

表 6：考核要点和评分标准

任务	考查点	描述	评分标准	分值（分）
Hadoop 平台及组件的部署管理	组件、配置、管理	按照 Hadoop 大数据平台部署要求完成平台安装部署和基本配置，按要求完成平台相关组件的部署与管理，以及集群节点的动态增删和高可用部署。检查平台进程运行状态，确定配置正确。	主要评分点包括 Linux 系统基本操作、Hadoop 平台安装、相关配置文件修改、平台相关组件部署、集群节点的动态增删、平台高可用部署。	15
数据采集	数据采集代码编写	按照要求完成数据采集代码的编写，使得数据采集程序能够正常运行，将采集到的数据保存在指定格式的路径下。	主要评分点包括数据采集请求构建、数据解析、数据存储、数据文件操作。	20

数据清洗与分析	数据清洗与分析代码编写	完成对数据的查询、整理和计算。进行编译、打包、发布，并在服务器上执行程序完成数据清洗工作，将清洗后的数据放置在指定路径下。将清洗后的数据加载到数据仓库后，根据项目需求使用命令或代码，完成数据分析查询，并将查询结果导出为数据文件。	主要评分点包括数据处理代码编写、数据文件解析、构建数据输出格式、打包发布、数据清洗执行。数据仓库建库、建表、查询等处理。	25
数据可视化	数据可视化代码编写	通过编写后台数据访问代码完成数据可视化后台开发，编写前端 Web 界面实现数据分析结果可视化呈现。	主要评分点包括可视化后台代码开发、可视化前端代码开发、前端展示。	20
综合分析	文档编写	根据项目要求，以数据可视化结果为依据，得出数据分析结论，生成分析报告并提交。	主要评分点包括能够按照赛项要求编写结论，能够按照要求提出正确的建议。	15
职业素养	职业素养	团队分工明确合理、操作规范、文明竞赛	主要评分点包括：竞赛团队分工明确合理、操作规范、文明竞赛。	5

（三）成绩审核及公布方法

最终以省上下发文件为准执行。

十、申诉与仲裁

最终以省上下发文件为准执行。

十一、竞赛观摩

由于赛项特点及赛场条件限制，本赛项不设置现场观摩环节。

十二、竞赛视频

赛场内部署无盲点录像设备，能实时录制赛场情况。

十三、竞赛须知

（一）参赛队须知

1. 参赛队按照大赛赛程安排，凭赛项执委会颁发的参赛证和有效身份证件参加比赛及相关活动。

2. 赛项执委会统一安排各参赛队在比赛前一天进入赛场熟悉环境和设施情况。

3. 参赛队选手、领队和指导教师要有良好的职业道德，严格遵守比赛规则和比赛纪律，服从裁判，尊重裁判和赛场工作人员，自觉维护赛场秩序。

4. 领队应负责赛事活动期间本队所有选手的人身及财产安全，如发现意外事故，应及时向赛项执委会报告。

5. 各学校组织代表队时，须为参赛选手购买大赛期间的人身意外伤害保险。

（二）领队和指导教师须知

1. 严格遵守赛场的各项规定，服从裁判，文明竞赛。如发现弄虚作假者，取消参赛资格，名次无效。

2. 领队和指导教师务必带好有效身份证件，在活动过程中佩戴“指导教师证”参加竞赛相关活动。

3. 各代表队领队要坚决执行竞赛的各项规定，加强对参赛人员的管理，做好赛前准备工作，督促选手带好证件等竞赛相关材料。

4. 在比赛期间要严格遵守比赛规则，不得私自接触裁判人员。

5. 竞赛过程中，未经裁判许可，领队、指导教师及其他人员一

律不得进入竞赛现场。

6. 领队和指导老师应及时查看有关赛项的通知和内容，认真研究和掌握本赛项竞赛的规程、技术规范和赛场要求，指导选手做好赛前的一切技术准备和竞赛准备。

（三）参赛选手须知

1. 参赛选手应严格遵守赛场规章、操作规程和工艺准则，保证人身及设备安全，接受裁判员的监督和警示，文明竞赛。

2. 参赛选手应按照规定时间抵达赛场，凭身份证、统一发放的参赛证（两证必须齐全），完成入场检录、抽签确定竞赛工位号，不得迟到早退。

3. 参赛选手凭竞赛工位号进入赛场，不允许携带任何电子设备及其他资料、用品。

4. 参赛选手应在规定的时间段进入赛场，认真核对竞赛工位号，在指定位置就座。

5. 参赛选手入场后，迅速确认竞赛设备状况，填写相关确认文件，并确认签字（竞赛工位号）。

6. 参赛选手在收到开赛信号前不得启动操作。在竞赛过程中，确因计算机软件或硬件故障，致使操作无法继续的，经项目裁判长确认，予以启用备用计算机。

7. 赛项任务书及相关资料，均保存在竞赛环境的“大赛资料”中。参赛选手应在竞赛规定时间内完成任务书内容，并按照要求，将相应文档上拷贝到U盘。

8. 参赛选手需及时保存工作记录。对于因各种原因造成的数据丢失，由参赛选手自行负责。

9. 参赛队所提交的答卷采用竞赛工位号进行标识，不得出现地

名、校名、姓名、参赛证编号等信息，否则取消竞赛成绩。

10. 竞赛过程中，因严重操作失误或安全事故不能进行比赛的（例如因操作原因发生短路导致赛场断电的、造成设备不能正常工作的），现场裁判员有权中止该队比赛。

11. 在比赛中如遇非人为因素造成的设备故障，经裁判确认后，可向裁判长申请补足排除故障的时间。

12. 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。

13. 竞赛操作结束后，参赛选手需要根据任务书要求，将相关成果文件拷贝至U盘，填写结束比赛相关确认文件，并签字确认（竞赛工位号）。因参赛选手未能按要求，将相应的文档等拷贝至U盘的，竞赛成绩计为零分。

14. 竞赛时间结束，选手应全体起立，停止操作。将资料和工具整齐摆放在操作平台上，经工作人员清点后可离开赛场，离开赛场时不得带走任何资料。

15. 在竞赛期间，未经执委会批准，参赛选手不得接受其他单位和个人进行的与竞赛内容相关的采访。参赛选手不得将竞赛的相关信息私自公布。

16. 符合下列情形之一的参赛选手，经裁判组裁定后中止其竞赛：

（1）不服从裁判员/监考员管理、扰乱赛场秩序、干扰其他参赛选手比赛，裁判员应提出警告，二次警告后无效，或情节特别严重，造成竞赛中止的，经裁判长确认，中止比赛，并取消竞赛资格和竞赛成绩。

（2）竞赛过程中，由于选手人为造成计算机、仪器设备及工具

等严重损坏，负责赔偿其损失，并由裁判组裁定其竞赛结束与否、是否保留竞赛资格、是否累计其有效竞赛成绩。

(3) 竞赛过程中，产生重大安全事故、或有产生重大安全隐患，经裁判员提示没有采取措施的，裁判员可暂停其竞赛，并由裁判组裁定其竞赛结束与否、是否保留竞赛资格、是否累计其有效竞赛成绩。

(四) 工作人员须知

1. 赛场工作人员由赛项执委会统一聘用并进行工作分工，进入竞赛现场须佩戴赛项执委会统一提供的胸牌。

2. 赛场工作人员需服从赛项执委会的管理，严格执行赛项各项比赛规则，执行各项工作安排，积极维护好赛场秩序，坚守岗位，为赛场提供有序的服务。

3. 赛场工作人员进入现场，不得携带任何通讯工具或与竞赛无关的物品。

4. 参赛队进入赛场，现场裁判应按规定审查参赛选手带入赛场的物品，如发现不允许带入赛场的物品，交由参赛队随行人员保管，赛场不提供保管服务。

5. 赛场工作人员在竞赛过程中不回答选手提出的任何有关比赛技术问题，如遇争议问题，应及时报告裁判长。

附件一：大数据技术与应用赛项竞赛试题（样卷）

近年来随着 IT 产业的加速发展，全国各地对 IT 类的人才需求也越来越多“ABC 公司”为了明确今后 IT 产业人才培养方向，在多地
进行 IT 公司岗位情况调研分析。你所在的小组将承担模拟调研分析的任务，通过在招聘网站进行招聘信息的爬取，获取到公司名称、工作地点、岗位名称、招聘要求、招聘人数等信息，并通过对数据的清洗和分析，得出各地域招聘人数，“大数据”相关职位招聘数量，以绘制雷达图展示各地平均薪资情况。

为完成该项任务，你所在的小组计划选用在业界广泛应用的“Python 和 JAVA”语言，作为整个项目的基础语言，并综合利用 requests 模块、MapReduce、MySQL、Flask 开源框架、Jinja2 模板引擎和 ECharts 组件提高开发效率并实现项目要求，由于本次为模拟任务，总数据量不会过大，项目组计划使用分布式节点 Hadoop 模式，本次项目环境搭建采用服务器集群方式，配置了小规模的技术演示环境，通过在招聘网站上爬取到的相关信息，使用 requests 模块、Hive、Python、JAVA 等手段对数据进行爬取、清洗、整理、计算、表达、分析，力求实现对 IT 人才就业信息拥有更清晰的掌握。

请按照下面步骤完成本次技术展示任务，并提交技术报告。

任务一：Hadoop 相关组件安装部署（15 分）

当前环境中已安装 Hadoop 运行环境和 MySQL 数据库，相关安装信息如下表所示，请在此环境基础上按照相关操作步骤安装 Hive 组件。

1. 将指定路径下的 Hive 安装包解压并更名；
2. 设置 Hive 环境变量；
3. 编辑 Hive 相关配置文件；
4. 初始化 Hive 元数据；
5. 启动并保存输出结果。

任务二：数据采集与数据预处理（20 分）

1. 从指定招聘网站中抓取数据，提取有效数据项，并保存为 json 格式文件；
2. 设置 post 请求参数并将信息返回给变量 response；
3. 将提取数据转化成 json 格式，并赋值变量；
4. 用 with 函数创建 json 文件，通过 json 方法，写入 json 数据；
5. 爬取的数据需要导入 hadoop 平台进行数据清洗与分析，在 HDFS 文件系统中创建文件夹，并将 json 文件上传到该文件夹下。

任务三：数据清洗与分析（25 分）

1. 为便于数据分析与可视化，需要对爬取出的数据进行清洗，使用 Java 语言编写数据清洗的 MapReduce 程序；
2. 将清洗程序上传至 hadoop，并对 HDFS 的原始数据进行清洗；
3. 将清洗后的数据加载到 Hive 数据仓库中；
4. 通过运行 HQL 命令完成数据分析统计；
5. 在 hive 中执行 sql 脚本，并查看表中大数据核心技能的出现次数。

任务四：数据可视化（20 分）

为更好的将数据分析结果表达出来，需要对数据分析的结束进行可视化呈现，可视化呈现，本次数据可视化需要呈现三部分内容：

1. 按要求使用柱状图展示各城市招聘人数，并在前端显示。要求：

主标题：各地域招聘人数

副标题：（--招聘人数变化趋势）

横坐标：城市信息，纵坐标：招聘人数

输出柱状图

2. 按要求使用折线图展示“大数据”相关职位招聘数量差异，并在前端显示。要求：

主标题：大数据相关职位分析

副标题：（--招聘数量变化趋势）

横坐标：岗位名称，纵坐标：岗位数量

输出折线图

3. 通过雷达图展示各地平均薪资的情况。要求：

主标题：各地平均薪资

输出雷达图

任务五：完成分析报告（15 分）

请结合数据分析结果回答以下问题：

1. 根据分析结果说明大数据岗位所需要的主要技能包含哪些，为什么（4 分）

2. 根据分析结果说明各地大数据产业发展情况（4分）
3. 根据市场需求分析，大数据行业的人才培养方向有哪些，为什么（4分）
4. 请简述，今后大数据产业地域发展方向在哪里（3分）
5. 竞赛结果提交要求：

（一）文档要求

任务成果需拷贝至提供的U盘中。在U盘中以XX工位号建一个文件夹（例如01），将所有任务成果文档保存至该文件夹中。

（二）文档要求

竞赛提交的所有文档中不能出现参赛队信息和参赛选手信息，竞赛文档需要填写参赛队信息时以工位号代替（XX代表工位号）。