2021年河北省省高职院校技能大赛
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

赛项归属：电子信息大类

二、竞赛目的

为贯彻落实国务院发布的《促进大数据发展行动纲要》和工业和信息化部发布的《大数据产业发展规划（2016-2020年）》。加快实施国家大数据战略，推动大数据产业健康快速发展，针对高职“大数据技术与应用”专业建设和发展的需求，通过引入大数据各个环节的实际应用场景，全面考察高职学生大数据技术基础、软件开发相关技术、Hadoop及其生态组件部署与管理、数据采集、数据清洗、数据分析和数据可视化等前沿的知识、技术技能以及职业素养和团队协作能力。

赛项围绕大数据产业各个岗位的实际需求和要求进行设计，通过大赛搭建校企合作的平台，深化产教融合，推进产教融合人才培养模式，提升大数据技术与应用专业及其他相关专业毕业生能力，推动院校和企业联合培养大数据人才，加强学校教育与产业发展的有效衔接，促进职业院校信息类相关专业共同发展，为国家战略规划提供大数据领域的高素质技能型人才。

三、竞赛内容

（一）选手需具备能力

本赛项基于企业真实项目和工作任务，结合企业岗位对学生职业技能的最新需求，在规定的时间内完成指定任务。其中，主要考核参赛选手在大数据平台部署管理、数据采集、数据清洗分析、数据可视化及综合分析等方面技能。此外，竞赛同时考核参赛选手工作组织和自我管理能力、沟通和人际交往能力、解决问题能力以及致力于紧跟行业发展步伐的自我学习能力。

本项目竞赛内容通过对技能实操表现来评估知识理解以及技能的熟练程度，将不再另外举行知识及理解性质的理论测试。

（二）竞赛模块

大数据技术与应用赛项基于企业真实项目，结合企业岗位技能需求，在 4 小时完成指定任务。

1. 竞赛内容

本竞赛结合国内行业、企业的实际业务和世界技能大赛标准来组织命题；本竞赛只考核技能部分，不涉及理论。本竞赛进行的技能实操考核，涉及 Hadoop 平台及组件的部署管理、数据采集、数据清洗与分析、数据可视化、综合分析。

模块编号	模块名称	竞赛时间	分数		
			评价分	测量分（%）	合计（%）
A	Hadoop 平台及组件的部署管理	4 小时	/	15	15
B	数据采集		/	20	20
C	数据清洗与分析		/	25	25

D	数据可视化		/	20	20
E	综合分析		/	20	20
总计					100

2.模块介绍

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，使用 Java、Python、Scala 等开发语言，完成数据清洗、数据存储、数据转化、数据分析、数据预测及数据推送等一系列数据操作
	多表数据合并和离群值处理	
	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 及 Web 前端等编程语言，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

（一）选手构成

本赛项为团体技能赛，每支参赛队由 3 名选手组成，必须为河北省高等学校在籍高职高专类学生。其中，参赛选手年龄须不超过 25 周岁(年龄计算的截止时间以 2021 年 4 月 17 日为准)，其性别和年级不限。指导教师须为本校专兼职教师，每支参赛队限报 2 名指导教师。

（二）竞赛时间安排

本赛项分五个模块。所有参赛选手在指定时间、按照比赛要求完成比赛任务。竞赛时间为 4 小时。

五、竞赛流程

（一）竞赛流程图

2021 年大数据技术与应用赛项的竞赛流程如图 1 所示。

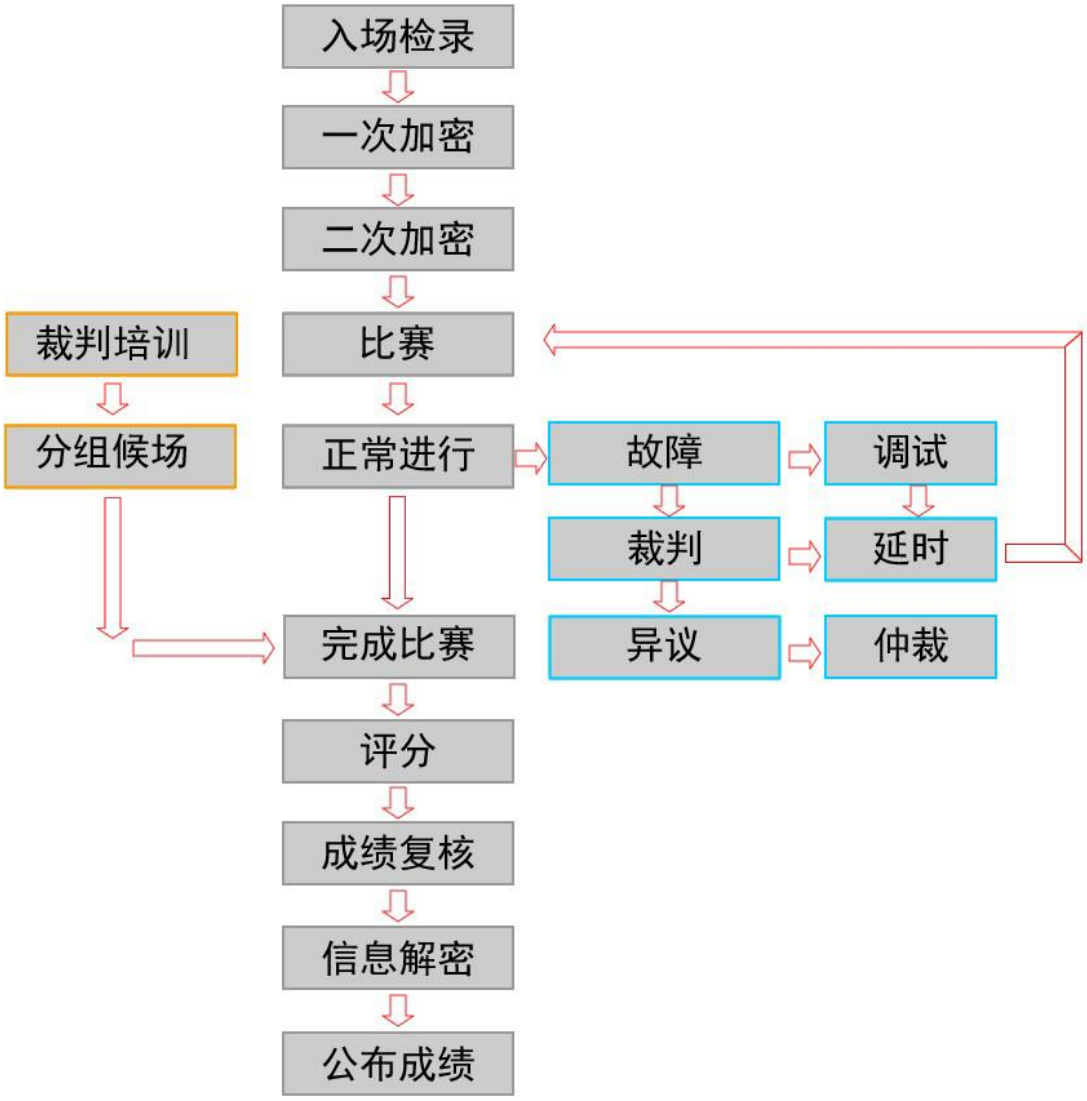

图 1　竞赛流程图

（二）竞赛时间安排

日期	时　间	内　容
第一日	8:00-14:00	报到　（地点：科工大西校区 B 区大厅）
	15:00-16:00	领队会、赛前说明　（地点：科工大西校区 B 区 112 室）
	16:00-16:30	选手熟悉赛场
第二日	7:30-8:20	领队抽取场次签及检录顺序号
	8:20-8:35	裁判长在监督长监督下，考试题目发布
	8:35-8:50	赛场检录；竞赛选手抽签、确定竞赛工位号
	8:50-9:00	竞赛选手进入赛位，检查赛位设备； 现场裁判发放题目、宣布竞赛注意事项；

	9:00-13:00	竞赛选手完成竞赛任务
	13:30-20:00	竞赛成绩评定，公布竞赛成绩

六．竞赛赛卷

比赛当天由裁判长公布比赛赛题，裁判长在监督长监督下，开启竞赛试题。

比赛完成后，包括参赛选手在内的任何人，都不得将赛题带离赛场，由现场裁判对赛题进行回收。

赛卷具体参考样卷见附件。

七、竞赛规则

1. 参赛队及参赛选手资格。参赛选手须为高职院校全日制在籍注册学生、本科院校中高职类全日制在籍注册学生、五年制高职四、五年级在籍注册学生。参赛选手年龄须不超过 25 周岁(年龄计算的截止时间以 2021 年 4 月 16 日为准)。凡在往届全国职业院校技能大赛中获本赛项高职组一等奖的选手，不能再报名参赛。

2. 比赛工位通过抽签决定，比赛期间参赛选手原则上不得离开比赛场地。

3. 竞赛所需的硬件、软件和辅助工具统一提供，参赛队不得使用自带的任何具有存储和通讯功能的设备，如硬盘、光盘、U 盘、手机、随身听、智能手表、PDA 等。

4. 参赛选手在赛前 20 分钟，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。

5. 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。

6. 比赛时间结束，选手应全体起立，结束操作。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。

7. 赛项裁判应严格遵守赛项各项规章制度，确保比赛公平、公正、公开。比赛当天 8:00 起，赛项裁判应上交所有通信设备，由赛项执委会统一保管，并安排赛项裁判在指定区域休息或工作，直至赛项成绩评定结束。

8. 比赛结束，经加密裁判对各参赛选手提交的竞赛结果进行第三次加密后，评分裁判方可入场进行成绩评判。

最终竞赛成绩经复核无误，由裁判长、监督长签字确认后，以纸质形式向全体参赛队进行公布，并在闭赛式上予以宣布。

9. 本赛项各参赛队最终成绩，由承办单位信息员录入赛务管理系统。承办单位信息员对成绩数据审核后，将赛务系统中录入的成绩导出打印，经赛项裁判长审核无误后，签字。

承办单位信息员将裁判长确认的电子版赛项成绩上传赛务管理系统；同时，将裁判长签字的纸质打印成绩单报送大赛执委会。

10. 赛项结束后，专家工作组根据裁判判分情况，分析参赛选手在比赛过程中对各知识点、技术的掌握程度，并将分析报告报备大赛执委会办公室，执委会办公室根据实际情况适时公布。

11. 赛项中每个比赛环节裁判判分的原始材料和最终成绩等结果性材料，经监督组人员和裁判长签字后，装袋密封留档；并由赛项承办院校封存，委派专人妥善保管。

七、竞赛环境

（一）赛场布局要求

竞赛场地包括参赛选手竞赛区域、展示平台区域、裁判区域、设备耗材区、技术支持区、服务区。

1. 参赛选手竞赛区域。每个竞赛工位标有醒目的工位编号，确保参赛队之间互不干扰。赛场要求竞赛过程全程无死角视频监控，监控录像保存 3 个月。环境标准要求保证赛场采光（大于 500 lux）、照明和通风良好；提供稳定的水、电，并提供应急的备用电源；提供足够的干粉灭火器材。

2. 休息区域需要与比赛场地分开，供参赛队领队、指导教师及工作人员休息，并开展其他相关活动。

3. 裁判区域。供裁判休息及工作场地。共配有电脑 10 台；A4 激光打印机 1 台；桌椅 10 套；饮水机、纸杯、文具用品若干。

4.技术支持区。为技术支持人员的工作场地，为参赛选手竞赛提供技术支持。

5. 服务区。提供医疗等服务保障，并用隔离带隔离。

（二）赛场选手安全防护要求

1. 参赛选手应严格遵守设备安全操作规程。

2. 参赛选手停止操作时，应保证设备的正常运行，比赛结束后，所有设备保持运行状态，不要拆、动硬件连接，确保设备正常运行，实现正常评分。

3. 参赛选手应遵从安全规范操作，例如：ESD(静电放电)设备安全使用及储存。

4. 参赛选手应保证设备和信息的完整及安全。

（三）赛事安全要求

1. 禁止选手及所有参加赛事的人员，携带任何有毒有害物品进入竞赛现场。

2. 承办单位应设置专门的安全防卫组，负责竞赛期间健康和安全事务。主要包括检查竞赛场地、与会人员居住地、车辆交通及其周围环境的安全防卫；制定紧急应对方案；监督与会人员食品安全与卫生；分析和处理安全突发事件等工作。

3. 赛场须配备相应医疗人员和急救人员，并备有相应急救设施。

（四）赛事开放要求

1. 赛场内除指定的裁判、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内。

2. 允许进入赛场的人员，不得使用录像设备长时间拍摄选手工位、屏幕。

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、喧哗。

5.经组委会允许的赞助商和负责宣传的媒体记者，按竞赛规则的要求进入赛场相关区域。

上述相关人员不得妨碍、干扰选手竞赛，不得有任何影响竞赛公平、公正的行为。

（五）赛事绿色环保要求

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。

3. 赛场设置排烟除尘系统，尽可能地减少和控制烟尘。

八、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。

（一）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

（二）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

九、技术平台

（一）竞赛设备

比赛器材、技术平台：新华三大数据竞赛管理系统（合作企业：新华三技术有限公司；品牌：H3C），技术平台软硬件设备组成如下：

序号	设备名称	数量	备注
1	服务器	1	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*2.3GHz 2、内存模块：8*32GB 3、硬盘模块：6*600GB SAS 10K 4、网口：4 端口千兆电接口网卡-360T-B2 5、1+1 冗余电源
2	大数据竞赛平台（H3C-ADEkvm-DT）	1	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。
3	PC 机	3	竞赛选手比赛使用。性能相当于 i5 处理器，8G 以上内存，1TB 以上硬盘，显示器要求 1024*768 以上。
4	交换机	1	1.机架式交换机 2.端口：≥24 个 10/100/1000Base-TX 以太网端口； 3.速度：10/100/1000Base； 4.全千兆三层交换机，支持访问控制。

备注：实际赛场需要的服务器、PC 机和交换机数量取决于参赛队伍数量。

（二）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

		Hive 1.1.0
		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
	开发语言	Python 3.6 64bit
		Java 8
		Scala 11
	开发工具	Pycharm 2019 （Community Edition）
		IDEA 2019 （Community Edition）
	数据采集组件	Requests
		Scrapy
	数据可视化组件	ECharts 4.0.4
		Flask
		Jinja2
		Matplotlib
	文档编辑器	WPS2020 以上
	输入法	拼音输入法

十、成绩评定

（一）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，大数据技术与应用赛项评分采用赛项结果评分方法，始终贯彻落实大赛一贯坚持的公平、公正和公开原则。

赛项评分依据选手固化在实操任务中的成果，通过评分裁判对比赛成果再现的方法评分，并兼顾团队协作精神和职业素养综合评定。

参与大赛赛项成绩管理的组织机构包括裁判组、监督组和仲裁组等。裁判组实行“裁判长负责制”，设裁判长 1 名，全面负责赛项的裁判与管理工作。

裁判员根据比赛工作需要分为检录裁判、加密裁判、现场裁判和评分裁判。检录裁判负责对参赛队伍（选手）进行点名登记、身份核对等工作；加密裁判负责组织参赛队伍（选手）抽签并对参赛队伍（选手）的信息进行加密、解密；现场裁判按规定做好赛场记录，维护赛场纪律；评分裁判负责对参赛队伍（选手）的技能展示、操作规范和竞赛成果等按赛项评分标准进行评定。

监督组对裁判组的工作进行全程监督，并对竞赛成绩抽检复核。

仲裁组负责接受由参赛队领队提出的对裁判结果的申诉，组织复议并及时反馈复议结果。

（二）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、错误不传递、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。

（三）成绩审核方法

竞赛结束后，由裁判长向裁判员核实竞赛过程中有无异常。如无异常，成绩单由裁判长签字确认并封存直至公布成绩时开启。

如有异常，在裁判长主持下，由专家组成员、裁判员、仲裁员和监督员共同处理。

（四）成绩公布方法

竞赛成绩经复核无误后，经裁判长审核签字后，以赛项组委会最终公布结果为准

竞赛结束后，如参赛队对比赛成绩有异议，提出异议申诉或仲裁，可按照相关规定进行申诉和仲裁，按照仲裁结果公布竞赛成绩。

十一、赛项安全

赛事安全是技能竞赛一切工作顺利开展的先决条件，是赛事筹备和运行工作必须考虑的核心问题。赛项执委会采取切实有效措施保证大赛期间参赛选手、指导教师、裁判员、工作人员及观众的人身安全。

（一）比赛环境

1. 执委会须在赛前组织专人对比赛现场、住宿场所和交通保障进行考察，并对安全工作提出明确要求。赛场的布置，赛场内的器材、设备，应符合国家有关安全规定。如有必要，也可进行赛场仿真模拟测试，以发现可能出现的问题。承办单位赛前须按照执委会要求排除安全隐患。

2. 严格控制与参赛无关的易燃易爆以及各类危险品进入比赛场地，不许随便携带书包进入赛场。

3. 配备先进的仪器，防止有人利用电磁波干扰比赛秩序。大赛现场需对赛场进行网络安全控制，以免场内外信息交互，充分体现大赛的严肃、公平和公正性。

4. 大赛期间，承办单位须在赛场管理的关键岗位，增加力量，建立安全管理日志。

（二）生活条件

比赛期间，食宿自理。

（三）组队责任

1.各学校组织代表队时，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。赛项出现重大安全问题可以停赛，是否停赛由赛区执委会决定。事后，赛区执委会应向大赛执委会报告详细情况。

（五）处罚措施

1.因参赛队伍原因造成重大安全事故的，取消其获奖资格。

2.参赛队伍有发生重大安全事故隐患，经赛场工作人员提示、警告无效的，可取消其继续比赛的资格。

3.赛事工作人员违规的，按照相应的制度追究责任。情节恶劣并造成重大安全事故的，由司法机关追究相应法律责任。

十二、竞赛须知

（一）参赛队须知

1．参赛队名称：统一使用规定的学校代表队名称，不使用其他组织、团体的名称；

2．参赛队组成：每支参赛队由 3 名参赛选手组成，须为同校在籍学生，其中队长 1 名。每支参赛队可配 1-2 名指导教师，指导教师

须为本校专兼职教师。不接受跨校组队，同一学校可报名多支参赛队伍；

3．各参赛院校应指定1名负责人任赛项领队，全权负责该校参赛事务的组织、协调和领导工作。

4．参赛选手及指导教师在报名获得确认后，原则上不再更换。如在筹备过程中，参赛选手和指导教师因故不能参赛，须由其所在学校供职部门于赛项开赛前10个工作日之前出具书面说明，经大赛执委会办公室核实后予以更换。允许队员缺席比赛；允许指导教师缺席比赛。

5．参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。

6．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。

7．参赛队选手、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。

8．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。

9．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”、“领队证”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．如对竞赛过程有疑议，由领队和指导教师以书面形式向大赛仲裁委员会反映，但不得影响竞赛进行。

7．对申诉的仲裁结果，领队要带头服从和执行，并做好选手工作。参赛选手不得因申诉或对处理意见不服而停止竞赛，否则以弃权处理。

8．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、学生证，以及统一发放的参赛证，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，并由参赛队长确认签字（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，并由参赛队长签字确认（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、或有产生重大安全事故隐患，经裁判员提示没有采取措施的，裁判员可暂停其竞赛，由裁判组裁定其竞赛结束，保留竞赛资格和有效竞赛成绩。

（四）工作人员须知

1．竞赛现场设现场裁判组，裁判长 1 名，现场裁判若干名。裁判要秉公裁判，监督检查参赛队安全有序竞赛。如遇疑问或争议，须请示裁判长裁决，裁判长的决定为现场最终裁定。

2．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。

3．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。

4．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。

5．参赛队进入赛场，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。

6．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。

十三、申诉与仲裁

（一）申诉

1、参赛队对不符合竞赛规定的设备、工具、软件，有失公正的评判、奖励，以及对工作人员的违规行为等，均可提出申诉。

2、申诉应在竞赛结束后 1 小时内提出，超过时效不予受理。申诉时，应按照规定的程序由参赛队领队向赛项仲裁工作组递交书面申诉报告。报告应对申诉事件的现象、发生的时间、涉及到的人员、申诉依据与理由等进行充分、实事求是的叙述。事实依据不充分、仅凭主观臆断的申诉将不予受理。申诉报告须有申诉的参赛选手、领队签名。

3、赛项仲裁工作组收到申诉报告后，应根据申诉事由进行审查，3 小时内书面通知申诉方，告知申诉处理结果。

4、申诉人不得采取过激行为刁难、攻击工作人员，否则视为放弃申诉。

（二）仲裁

赛项设仲裁工作组接受由代表队领队提出的对裁判结果等方面问题的申诉。赛项仲裁工作组在接到申诉后的 2 小时内组织复议，并及时反馈复议结果。仲裁工作组的仲裁结果为最终结果。

十四、竞赛样题

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：

1）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。

2）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）。