第十二届河南省高职院校技能大赛
暨2019年全国职业院校技能大赛高职组河南省选拔赛
“大数据技术与应用”项目

竞赛规程

河南省“大数据技术与应用”项目大赛组委会

2019年4月01日

一、赛项名称

赛项编号：GZ-2019032

赛项名称：大数据技术与应用

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、竞赛方式

1. 比赛以团队方式进行,每个参赛队由 1 名领队(可由指导教师兼任)、2 名指导教师、3 名选手（其中队长 1 名）组成，指导教师须为本校专职教师。

2. 竞赛时间 4 小时。

四、竞赛内容

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，具体包括:

1. 掌握基于 Hadoop 离线分析平台，按照项目需求配置大数据组件并按照需求进行合理配置;

2. 掌握基于 Web 页面的数据采集相关技术，完成指定数据的采集及处理能力;

3. 综合利用 MapReduce 技术、分布式存储系统 HDFS、数据仓库 Hive 等工具及技术，使用 Java、Python 等开发语言，完成数据清洗、数据存储、数据转化、数据分析及数据推送等一系列大数据操作;

4. 综合运用 HTML、CSS、JavaScript 等开发语言，结合 Flask 前端框架、Jinja2 开源模板引擎、Echarts 数据可视化组件，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。

竞赛时间 4 小时，竞赛连续进行。竞赛内容构成如下:

考核环节	考核知识点和技能点
平台组件配置	Hive 基本配置
	Hive 参数设置
数据采集（Python）	使用工具（Chrome 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
	数据采集网络请求构建
	采集数据解析及关键数据提取
	本地目录操作、文件创建、读写
数据清洗（Java、Linux ）	HDFS 数据文件读取、解析、清洗过滤，分区
	MapReduce 程序的编译、打包、发布
	执行 MapReduce 程序，完成数据清洗
数据存储分析（Linux Shell）	Hive 建表
	Hive 数据加载
	HQL 编写、数据查询统计
	Sqoop 数据推送

数据可视化（Python、HTML、CSS、JavaScript）	网页后台代码编写
	基于 Flask 前端框架、Jinja2 开源模板引擎、Echarts 数据可视化组件实现可视化渲染编码
综合分析	文档能力、数据分析能力

竞赛各阶段分值权重和时间分布如下:

阶段	竞赛时间	分值权重
大数据组件配置	4 小时	权重 10%
数据采集		权重 15%
数据清洗		权重 20%
数据存储分析		权重 20%
数据可视化		权重 20%
数据分析报告		权重 10%
团队分工明确合理、操作规范、文明竞赛		权重 5%

五、竞赛流程

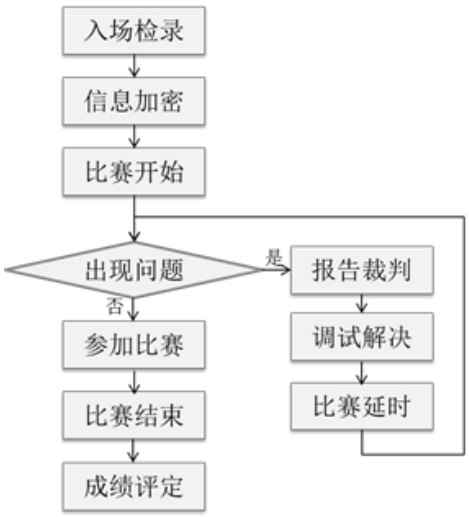
根据竞赛任务要求，参赛队伍在 4 小时竞赛时间内须完成竞赛任务，参赛队伍须按顺序完成各项任务，但每项任务用时可自行掌握。

（一）竞赛时间安排

日期	时　间	内　容	地点
4 月 10 日	14:00-16:00	参赛选手报到，发材料、参赛证等	住宿酒店
	16:30-17:00	领队抽取顺序号（第一次抽号），确定上、下午场次	实训楼 4402
	17:00-17:30	选手熟悉赛场	实训楼 4405C
4 月 11 日上午	7:30	上午场：检录入场；选手抽工位号（第二次抽号）	实训楼 4402
	8:00-12:00	上午场：比赛	实训楼 4405C
	12:00-12:10	上午场：赛后设备及提交资料确认	
	12:10-13:00	上午场：选手离场封闭	实训楼 4405B3
	12:10-13:30	裁判员成绩回收及恢复设备	

4 月 11 日 下午	12:50	下午场：选手集合进行封闭	实训楼 4402
	13:30-13:50	下午场：检录入场；选手抽工位号（第二次抽号）	实训楼 4402
	14:00-18:00	下午场：比赛	实训楼 4405C
	18:00-18:10	下午场：赛后设备及提交资料确认	
	18:10	清场、开始统计成绩	

（二）竞赛流程



六、竞赛试题

本赛项样题详见《附件一》。

七、竞赛规则

1. 比赛时间为 4 个小时，比赛过程连续进行。

2. 参赛队的竞赛工位号采用抽签方式确定。赛题以任务书形式发放，竞赛参考资料在赛前植入参赛选手的计算机，参赛队根据任务书要求完成竞赛任务。

3. 组委会统一布置竞赛需要的软硬件环境。选手不得私自携带任何移动存储、辅助工具、移动通信等进入赛场。

4. 参赛选手报到当天可预先熟悉比赛场地，但不得进行现场练习。参赛选手按规定时间到达指定地点，凭参赛证、学生证和身份证（三证必须

齐全）进入赛场。选手迟到 10 分钟取消比赛资格。

5. 参赛选手不得携带通讯工具和其它未经允许的资料、物品进入比赛场地，不得中途退场。如出现较严重的违规、违纪、舞弊等现象，经裁判组裁定取消比赛成绩。

6. 在竞赛过程中，参赛选手如有疑问，应举手示意，考场裁判长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，考场裁判长、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，经考场裁判长确认，予以启用备用设备。

7. 比赛过程中，参赛选手须严格遵守操作标准和规范，保证自身安全，并接受裁判员的监督和警示；若因设备故障导致选手中断或终止比赛，由大赛裁判长视具体情况做出裁决。

8. 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。

9. 现场比赛结束，经裁判员确认后方可离开赛场。

10. 各赛项由裁判员现场评分，经裁判长签字确认后上交组委会。

11. 每个参赛队必须参加所有专项的比赛。参赛选手应严格遵守赛场纪律，服从指挥，着装整洁，仪表端庄，讲文明礼貌。各地代表队之间应团结、友好、协作，避免发生任何形式的矛盾。

12. 其它未尽事宜，将在赛前向各领队做详细说明。

八、技术规范

本赛项的技术规范包括：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等。

（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

设备类别	数量	设备用途	基本配置
竞赛服务器	7 台。 采用集群管理方式；2 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)
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，1TB 以上硬盘，显示器要求 1024*768 以上

（二）软件环境

设备类型	软件类别	软件名称、版本号
竞赛服务器集群	竞赛环境大数据集群操作系统	CentOS 7.4.1708 mini
	大数据平台组件	Hadoop 2.6.0
		Hive 1.1.0
		Sqoop 1.4.7
		Scala 2.11.8
开发客户端	PC 操作系统	Win 10 64 位
	浏览器	Chrome
	终端模拟软件	XShell
	开发语言	Python 3.6.4 64bit
		Java 8
	开发工具	Pycharm 2019.1（Community Edition）
		IDEA 2019.1（Community Edition）
	数据采集组件	Requests 2.19.1
	数据可视化组件	Flask 0.12.2
		Jinja2 2.10
		ECharts 4.1.0
	文档编辑器	Office 2007 以上
	输入法	搜狗拼音输入法

十、成绩评定

（一）评分标准制定原则

严格遵守公平、公正的原则，采用赛项结果评分方法。赛项评分依据选手固化在实操任务中的成果，通过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。

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、累计总分的计分方式。

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）评分细则

任务	考查点	描述	评分标准	分值
环境配置	组件与配置	在已部署完成的大数据离线分析平台的设备上完成数据仓库组件 Hive 的配置，运行测试程序，确定配置正确。	主要评分点包括更改文件名、Hive 环境变量设置、连接 MySQL 数据库、初始化 Hive。	10
数据采集	数据采集代码编写	按照要求完成特定函数的编写，使得数据采集程序能够正常运行，将采集到的数据保存在文件中。	主要评分点包括数据请求构建、数据解析、数据存储、数据文件操作。	15
数据清洗	数据清洗代码编写	使用 Java 语言完成 MapReduce 程序的代码编写、程序打包发布并在服务器运行完成数据清洗工作，将清洗后的数据放置在指定路径下	主要评分点包括数据处理代码编写、Json 数据解析、构建数据输出格式、打包发布、数据清洗执行。	20
数据分析	数据分析代码编写	将清洗后的数据加载到 Hive 数据仓库中后，根据项目需求使用 HQL 语句，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 Hive 建库、Hive 建表、HQL 查询。	20
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，使用 Flask 前端框架、Jinja2 开源模板引擎或 Echarts 完成数据可视化。	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

数据分析报告	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。	10
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

（四）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。

（五）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

1.比赛期间，原则上由执委会统一安排参赛选手和指导教师食宿。承办单位须尊重少数民族的信仰及文化，根据国家相关的民族政策，安排好少数民族选手和教师的饮食起居。

2.比赛期间安排的住宿地应具有宾馆/住宿经营许可资质。

3.各赛项的安全管理，除了可以采取必要的安全隔离措施外，应严格遵守国家相关法律法规，保护个人隐私和人身自由。

（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. 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。

3. 参赛选手及指导教师在报名获得确认后，原则上不再更换。允许队员缺席比赛；允许指导教师缺席比赛。

4. 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. 如对竞赛过程有疑议，由领队和指导教师负责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. 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. 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

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，2 小时内书面通知申诉方，告知申诉处理结果。仲裁工作组的仲裁结果为最终结果。

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。

附件一：大数据技术与应用赛项竞赛试题（样卷）

一、 竞赛时间、内容及总成绩

（一）竞赛时间

竞赛时间共为 4 小时，参赛队自行安排任务进度，休息、饮水、如厕等不设专门用时，统一含在竞赛时间内。

（二）竞赛内容概述

序号	任务名称	具体内容
任务一	组件配置	按照大数据离线分析平台需求，需要完成数据仓库组件的配置。
任务二	数据采集	编写数据采集代码对目标电子商务网站的数据进行采集，目标电子商务网站进行数据采集。
任务三	数据清洗	对采集到的数据进行不合规数据的清洗工作。请完成 MapReduce 程序的代码编写、程序打包发布并在服务器运行完成数据清洗工作。
任务四	数据存储与分析	将清洗后的数据加载到 Hive 数据仓库中后按照需求对数据进行分析。
任务五	数据可视化	编写前端数据可视化代码，完成数据可视化操作
任务六	编写数据分析报告	根据要求编写数据分析报告

（三）竞赛总成绩

“大数据技术与应用”赛项竞赛总成绩为 100 分，其中包含赛场职业素养 5 分。

二、 任务须知

1. 每组同学分配一台竞赛服务器、三台客户机，拥有独立 IP 组。
2. 本次比赛采用统一网络环境比赛，请不要随意更改客户端的网络地址信息，对于更改客户端信息造成的问题，由参赛选手自行承担比赛损失；
3. 请不要恶意破坏竞赛环境，对于恶意破坏竞赛环境的参赛者，组委会根据其行为予以处罚直至取消比赛资格。
4. 比赛过程中及时保存相关文档。
5. 比赛相关文档中不能出现参赛学校名称和参赛选手名称，以赛位号（工位号）代替。
6. 参赛选手请勿删除模板内容，若因删除导致任何问题后果自负。
7. 若同一文档由不同选手完成，须将文档合并后作为最终结果提交到 U 盘中。
8. 比赛中出现各种问题及时向监考裁判举手示意，不要影响其他参赛队比赛。

三、 任务说明

本项目要求完成手机销售数据分析，完成平台搭建、数据采集、数据清洗、数据存储分析、数据可视化及分析报告编写等工作。

提供的相关资源包括：

1. 数据，小于 1G
2. 数据仓库构建测试程序（内置在竞赛环境）
3. 数据采集代码模板（内置在竞赛环境）
4. 数据清洗代码模板（内置在竞赛环境）
5. 数据可视化代码模板（内置在竞赛环境）

6. 数据分析文档模板

任务一：环境安装配置

在已部署完成的大数据离线分析平台集群的指定设备上完成数据仓库组件的配置，运行测试程序，确定配置正确。

任务二：数据采集

按照要求完成特定函数的编写，使得数据采集程序能够正常运行，将采集到的数据保存在文件中。

任务三：数据清洗

使用 Java 语言完成 MapReduce 程序的代码编写、程序打包发布并在服务器运行完成数据清洗工作，将清洗后的数据放置在指定路径下。

任务四：数据存储分析

将清洗后的数据加载到 Hive 数据仓库中后，根据项目需求使用 HQL 语句，完成数据分析查询，并将查询结果导出为数据文件。

任务五：数据可视化

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，使用 Echarts 完成数据可视化。

任务六：编写数据分析报告

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。

四、竞赛结果提交要求

（一）提交方式

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）。