2023年云南省职业院校技能大赛
高职组“大数据应用开发”赛项规程

一、赛项名称

赛项名称：大数据应用开发

赛项组别：高职组

赛项归属产业：电子与信息大类

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，赛项以大数据应用开发为核心内容和工作基础，重点考查参赛选手基于Hadoop、Spark、Flink平台环境下，充分利用Spark Core、Spark SQL、Flume、Kafka、Flink、Hive、HBase、Redis、Maxwell、ClickHouse、MySQL等相关技术的特点，基于Scala、Java、JavaScript等开发语言，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，提升大数据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

赛项以大数据应用开发为核心内容和工作基础，重点考查参赛选手基于Hadoop、Spark、Flink平台环境下，充分利用Spark Core、Spark SQL、Flume、Kafka、Flink、Hive、HBase、Redis、Maxwell、ClickHouse、MySQL等技术的特点，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Hadoop平台、基于Spark的离线分析平台、基于Flink的实时分析平台，

在容器环境下，按照项目需求安装相关技术组件并按照需求进行合理配置；

2.掌握基于 Spark 的离线数据采集方式方法，完成指定数据的抽取并写入 Hive 分区表中。掌握基于 Flume、Maxwell 的实时数据采集，将数据写入 Kafka 中；

3.综合利用 Flink、Kafka、Hive、Redis、HBase、ClickHouse 等技术，使用 Java 开发语言，完成某电商系统的实时数据处理，包括使用 Flink 处理 Kafka 中的数据、实时数据仓库、将数据备份至 HBase 中、建立 Hive 外表、将数据处理结果存入 Redis、ClickHouse 中等操作；

4.综合利用 Spark、Hive、MySQL、HBase、ClickHouse 等相关技术，使用 Scala 开发语言，完成某电商系统的离线数据处理，包括 Hive 数据仓库、使用 Spark 处理离线数据、数据合并、去重、排序、数据类型转换、将数据处理结果存入 MySQL、HBase、ClickHouse 中等操作；

5.综合运用 HTML、CSS、JavaScript 等开发语言，Vue.js 前端技术，结合 ECharts 数据可视化组件，利用后端数据接口完成数据可视化；

6.根据竞赛过程，完成综合分析报告的编写；

7.竞赛时间 6 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
大数据平台环境搭建	Docker 基本操作
	Hadoop 完全分布式安装配置
	Spark 安装配置
	Flink 安装配置
	Hive 安装配置
	Kafka 安装配置

	Flume 安装配置
	ClickHouse 安装配置
	HBase 安装配置
数据采集	使用 Spark 抽取 MySQL 指定数据表中的增量数据到 ods 层的指定的分区表中
	使用 Flume 采集某端口的实时数据流并存入 Kafka 指定的 Topic 中
	使用 Maxwell 采集 MySQL 的 binlog 日志并存入 Kafka 指定的 Topic 中
实时数据处理	使用 Flink 消费 Kafka 中的数据并将数据分发至 Kafka 的 dwd 层中
	使用 Flink 消费 Kafka 中的数据的同时能够将数据备份至 HBase 中，同时建立 Hive 外表
	使用 Flink 对实时数据进行处理并将处理计算结果存入 Redis 中
	使用 Flink 对实时数据进行处理并将处理计算结果存入 ClickHouse 中
离线数据处理	使用 Spark 对 ods 层中的离线数据进行清洗，包括数据合并、去重、排序、数据类型转换等操作
	将清洗完的数据存入 dwd 层中
	根据 dwd 层的数据使用 Spark 对数据进行处理计算，并将计算结果存入 MySQL 中
	根据 dwd 层的数据使用 Spark 对数据进行处理计算，并将计算结果存入 HBase 中
	根据 dwd 层的数据使用 Spark 对数据进行处理计算，并将计算结果存入 ClickHouse 中
数据可视化	根据后端数据接口，基于 Vue.js、ECharts 的数据可视化编码（柱状图、折线图、饼状图等）
综合分析报告	文档能力、综合分析能力

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
大数据平台环境搭建	6 小时	权重 10%
数据采集		权重 15%
实时数据处理		权重 25%
离线数据处理		权重 20%
数据可视化		权重 15%
综合分析报告		权重 10%
团队分工明确合理、操作规范、文明竞赛		权重 5%

四、竞赛方式

1、比赛以师生联赛方式进行，不得跨校组队，同一学校的报名参赛队伍不超过 2 支。

2、每个参赛队由 1 名领队（可由参赛教师兼任）、4 名选手（1 名教师、3 名学生）组成，教师参赛选手须为职业院校教龄两年以上（含）的在职教师，学生参赛选手须为高等职业学校全日制在籍学生、五年一贯制四、五年级在籍学生、技工院校在籍学生，资格以报名时所具有的学生在校学籍及教师在职状态为准。参赛选手和指导教师报名获得确认后不得更换。

3、竞赛时间为 2023 年 9 月 26 日（报到），9 月 27 日（竞赛日）。竞赛时间 6 小时（09：00-15：00）。

五、竞赛流程

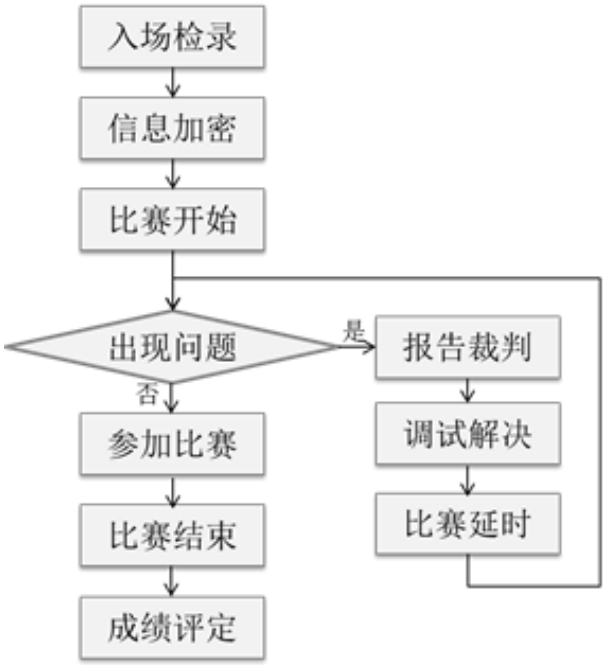
根据竞赛任务要求，参赛队伍在 6 小时竞赛时间内须完成竞赛任务，每项任务用时可自行掌握。

（一）竞赛时间安排

日期	时间	内容
9 月 26 日	14:00-18:00	各参赛队报到
	15:00-16:00	领队会、赛前说明
	16:00-16:30	选手熟悉赛场

9 月 27 日	8:00-9:00	赛场检录，竞赛选手进入赛位
	9:00-15:00	竞赛选手完成竞赛任务
	16:00-19:00	对选手提交的结果文件进行评分

（二）竞赛流程



六、竞赛命题

本赛项制定样题一套，并与本规程同步发布。具体详见附件 1。

七、竞赛规则

1.凡在往届省级职业院校技能大赛中获一等奖的学生，不能再参加同一项目同一组别的比赛。

2. 竞赛前 1 日安排各参赛队领队、参赛选手熟悉赛场。

3. 严禁参赛选手、赛项裁判、工作人员私自携带通讯、摄录设备进入比赛场地。

4. 参赛选手所需的硬件、软件和辅助工具统一提供，参赛队不得使用自带的任何有存储功能的设备，如硬盘、光盘、U 盘、手机、平板电脑等。

5. 所有参赛选手都必须携带身份证、学生证、参赛证（报名后由承办校统一发放）进行检录。

6. 参赛队在赛前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。

7. 比赛过程中，选手须严格遵守操作规程，确保人身及设备安全，并接受裁判员的监督和指示。因选手原因造成设备故障或损坏而无法继续比赛的，裁判长有权决定中止该队比赛；非因选手个人原因造成设备故障的，由裁判长视具体情况作出裁决。

8. 竞赛开始时统一发放本阶段赛卷，竞赛结束后，参赛选手要确认已成功提交竞赛要求的配置文件和文档，裁判员与参赛选手一起签字确认，参赛选手在确认后不得再进行任何操作。

9. 赛项成绩将于竞赛日当晚 24：00 前于 2023 年云南省职业院校技能大赛大数据应用开发赛项 QQ 群（群号：532871073）公示。

八、竞赛环境

1. 竞赛场地。竞赛场地分为：竞赛现场、裁判休息区。其中，竞赛现场又划分为：检录区、场内竞赛区、技术支持区。

2. 竞赛设备。场内竞赛区按照参赛队数量准备比赛所需的软硬件平台，为参赛队提供统一竞赛设备和备用设备。选手无需自带任何工具及附件。

3. 竞赛工位。竞赛现场各个工作区配备单相 220V/3A 以上交流电源。每个比赛工位上标明编号。

4. 技术支持区。为技术支持人员提供固定工位、电源保障。

5. 服务区。提供医疗等服务保障。

6. 竞赛场地应符合消防安全规定，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，赛场出入口专人负责。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。

九、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应用开发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

十、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	每支参赛队伍 1 台。根据参赛队数量，配备 10%的备份机器。	构建大数据平台集群	性能相当于 i5 处理器，64GB 以上内存，1TB 以上硬盘，网卡（千兆），显示器要求 1024*768 以上。
竞赛客户机	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，16GB 以上内存，1TB 以上硬盘，显示器要求 1024*768 以上。

（二）软件平台

由 2022 年全国职业院校技能大赛（高职组）大数据应用开发赛项合作企业——北京四合天地科技有限公司提供四合天地大数据实训管理系统。

（三）软件环境

设备类型	软件类别	软件名称、版本号
竞赛服务器	竞赛环境大数据集群操作系统	CentOS 7、Docker-CE 20.10
	大数据平台组件	Hadoop 3.1.3
		Hive 3.1.2
		HBase 2.2.3
		Spark 3.1.1
		Kafka 2.4.1
		Redis 6.2.6
		Flume 1.9.0
		Maxwell 1.29.0
		Flink 1.14.0
		ClickHouse 21.9.4
		JDK 1.8
		MySQL 5.7
开发客户端	PC 操作系统	Ubuntu18.04 64 位
	浏览器	Chrome
	开发语言	Scala 2.12
		Java 8
	开发工具	IDEA 2022 （Community Edition）
		Visual Studio Code 1.69
	数据库连接工具	MySQL Workbench
	SSH 工具	Asbru-cm 或 Ubuntu SSH 客户端
	API 测试工具	Postman API Platform
	数据可视化组件	Vue.js 3.0
		ECharts 5.1
	文档编辑器	WPS Linux 版

	输入法	搜狗拼音输入法 Linux 版

十一、成绩评定

（一）奖项设定

竞赛设参赛选手团体奖。奖项设置按参赛队数量确定，其中一等奖 10%，二等奖 20%，三等奖 30%。

（二）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，大数据应用开发赛项评分采用赛项结果评分方法，始终贯彻落实竞赛一贯坚持的公平、公正和公开原则。

参与竞赛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”。

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。

（三）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。

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，由裁判长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。

（四）评分标准

任务	考查点	描述	评分标准	分值（分）
大数据平	大数据相关平	在指定的宿主机上，基于 Docker 环境完成 Hadoop 完全分布式、Spark、	主要评分点包括 Hadoop 完全分布式安装配置、	10

台环境搭建（10分）	台组件安装配置	Flink、Hive、Kafka、Flume、ClickHouse、HBase 等的安装配置。	Spark 安装配置、Flink 安装配置、Hive 安装配置、Kafka 安装配置、Flume 安装配置、ClickHouse 安装配置、HBase 安装配置。	
数据采集（15分）	离线数据采集、实时数据采集	按照要求基于 Scala 语言完成特定函数的编写，使用 Spark 完成离线数据采集；按照要求使用 Linux 命令，利用 Flume、Maxwell、Kafka 等工具完成实时数据采集。	主要评分点包括 Spark 数据读取、数据存储、Flume 数据采集、Maxwell 数据采集、Kafka 等操作。	15
实时数据处理（25分）	实时数据处理计算代码编写	使用 Java 语言基于 Flink 完成 Kafka 中的数据消费，将数据分发至 Kafka 的 dwd 层中，并在 HBase 中进行备份同时建立 Hive 外表，基于 Flink 完成相关的数据指标计算并将计算结果存入 Redis、ClickHouse 中。	主要评分点包括 Flink 数据处理、数据指标计算、HBase、Hive、ClickHouse、Redis 等相关操作。	25
离线数据处理（20分）	离线数据处理计算代码编写	使用 Scala 语言基于 Spark 完成离线数据清洗、处理、计算，包括数据的合并、去重、排序、数据类型转换等并将计算结果存入 MySQL、HBase、ClickHouse 中。	主要评分点包括基于 Spark 的数据清洗、数据指标计算、HBase、Hive、ClickHouse、MySQL 等相关操作。	20
数据可视化（15分）	数据可视化代码编写	编写前端 Web 界面，调用后台数据接口，使用 Vue.js、ECharts 完成数据可视化。	主要评分点包括可视化前端代码开发、前端展示。	15
综合分析报告（10分）	文档编写	根据项目要求，完成综合分析报告编写。	主要评分点包括能够按照赛项要求进行综合分析。	10
职业素养（5分）	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

十二、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2小时内提出，超过时效将不予受理。申诉时，应按照规定的程序由参赛队领队向相应赛项裁判委员会递交书面申诉报告。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. 赛项裁判委员会收到申诉报告后，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。如受理申诉，要通知申诉方举办听证会的时间和地点；如不受理申诉，要说明理由。

4. 申诉人不得无故拒不接受处理结果，不允许采取过激行为刁难、攻击工作人员，否则视为放弃申诉。申诉人不满意赛项裁委会的处理结果的，可向赛项仲裁工作组提出复议申请。

（二）仲裁

1. 2023年云南省职业院校技能大赛（高职组）“大数据应用开发”赛项裁判委员会设仲裁工作组，负责受理竞赛中出现的申诉复议并进行仲裁，以保证竞赛的顺利进行和竞赛结果公平、公正。

2. 仲裁工作组的裁决为最终裁决，参赛队不得因对仲裁处理意见不服而停止比赛或滋事，否则按弃权处理。

附件一：大数据应用开发赛项竞赛试题（样卷）

一、 竞赛时间、内容及总成绩

（一）竞赛时间

竞赛时间共为6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	大数据平台环境搭建	按照任务书要求，需要基于 Docker 环境完成 Hadoop 完全分布式、Spark 安装配置、Flink 安装配置、Hive 安装配置、Kafka 安装配置、Flume 安装配置、ClickHouse 安装配置、HBase 安装配置等中的任意三个组件的安装配置
任务二	数据采集	按照任务书要求基于 Scala 语言基于 Spark 完成离线数据采集，将数据存入 Hive 的 ods 层中；按照要求使用 Linux 命令，利用 Flume、Maxwell、Kafka 等工具完成实时数据采集
任务三	实时数据处理	按照任务书要求使用 Java 语言基于 Flink 完成 Kafka 中的数据消费，将数据分发至 Kafka 的 dwd 层中，并在 HBase 中进行备份同时建立 Hive 外表，基于 Flink 完成相关的数据指标计算并将计算结果存入 Redis、ClickHouse 中
任务四	离线数据处理	按照任务书要求使用 Scala 语言基于 Spark 完成离线数据清洗、处理、计算，包括数据的合并、去重、排序、数据类型转换等并将计算结果存入 MySQL、HBase、ClickHouse 中
任务五	数据可视化	按照任务书要求编写前端代码，调用后台数据接口，使用 Vue.js、ECharts 完成数据可视化
任务六	综合分析报告	根据要求编写综合分析报告

（三）竞赛总成绩

“大数据应用开发”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. 每组参赛队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现场裁判举手示意，不要影响其他参赛队比赛。

三、 任务说明

本项目要求完成离线电商数据统计分析，完成大数据平台环境搭建、数据采集、实时数据处理、离线数据处理、数据可视化及综合分析报告编写等工作。

提供的相关资源包括：

1. 大数据环境搭建中需要用到的组件安装包
2. 电商相关脱敏业务数据
3. 大数据分析集群环境
4. 数据采集开发环境
5. 实时数据处理开发环境
6. 离线数据处理开发环境
7. 数据可视化开发环境
8. 综合分析报告文档模板

任务一：大数据平台环境搭建

按照任务书要求,需要基于 Docker 环境完成 Hadoop 完全分布式、Spark 安装配置、Flink 安装配置、Hive 安装配置、Kafka 安装配置、Flume 安装配置、ClickHouse 安装配置、HBase 安装配置等中的任意三个组件的安装配置。

任务二：数据采集

按照任务书要求基于 Scala 语言基于 Spark 完成离线数据采集，将数据存入 Hive 的 ods 层中；按照任务书要求使用 Linux 命令，利用 Flume、Maxwell 等工具完成实时数据采集，将数据存入 Kafka 指定的 Topic 中。

任务三：实时数据处理

按照任务书要求使用 Java 语言基于 Flink 完成 Kafka 中的数据消费，将数据分发至 Kafka 的 dwd 层中，并在 HBase 中进行备份同时建立 Hive 外表，基于 Flink 完成相关的数据指标计算并将计算结果存入 Redis、ClickHouse 中。

任务四：离线数据处理

按照任务书要求使用 Scala 语言基于 Spark 完成离线数据清洗、处理、计算，包括数据的合并、去重、排序、数据类型转换等并将计算结果存入 MySQL、HBase、ClickHouse 中。

任务五：数据可视化

按照任务书要求编写前端代码，调用后台数据接口，使用 Vue.js、ECharts 完成数据可视化。

任务六：综合分析报告

按照任务书要求，完成综合分析报告编写。

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）。