2023 年度“楚怡杯”湖南省职业院校技能竞赛
高职组“大数据技术与应用”赛项规程

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、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、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。

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 Scala 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 5 小时，竞赛连续进行。

四、竞赛方式

1、比赛以团队方式进行，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（可由指导教师兼任）、2 名指导教师、3 名选手（其中 1 队长 1 名）组成，指导教师须为本校专兼职教师，参赛选手和指导教师报名获得确认后不得随意更换。

五、竞赛时量

竞赛时长 300 分钟。

六、 名次确定办法

竞赛按照总成绩从高到低排序确定名次，不设并列名次。总成绩相同时，实时数据处理任务成绩高者名次排前，再相同者，依次按离线数据处理、数据采集、数据可视化、大数据平台环境搭建单项成绩高低确定排名。

七、 评分标准与评分细则

（一）评分标准

满分 100 分，总成绩为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、综合分析报告、职业素养得分之和。各部分分值权重见下表:

序号	阶段	分值权重
1	大数据平台环境搭建	权重 10%
2	数据采集	权重 15%
3	实时数据处理	权重 25%
4	离线数据处理	权重 20%
5	数据可视化	权重 15%
6	综合分析报告	权重 10%
7	职业素养	权重 5%

（二）评分细则

任务	考查点	描述	评分标准	分值（分）
大数据平台环境搭建（10分）	大数据相关平台组件安装配置	在指定的宿主机上，基于 Docker 环境完成 Hadoop 完全分布式、Spark、Flink、Hive、Kafka、Flume、ClickHouse、HBase 等的安装配置。	主要评分点包括 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。	10
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 Scala 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

八、赛项相关设施设备技术参数

（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 10 支参赛队伍共用 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	每队性能相当于 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。

U 盘	每支参赛队伍 1 个	提交成绩时使用	

（二）软件平台

依照国赛规程，根据赛点实际情况该赛项选用 2022 年全国职业院校技能大赛（高职组）大数据技术与应用赛项合作企业——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

	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

九、选手须知

（一）选手自带工（量）具及材料清单

参赛选手无需自带工具。

（二）主要技术规范及要求

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。

1.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

2.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

（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，经裁判长确认，予以启用备用计算机。

7. 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，须向现场裁判举手示意，经裁判长许可并完成记录后，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）竞赛直播

1.赛点提供全程无盲点录像。

2.可在赛点指定区域通过网络监控观摩比赛。

十、样题

见附件

附件一：大数据技术与应用赛项竞赛试题（样卷）

2023 年度“楚怡杯”湖南省职业院校技能竞赛

高职高专组电子与信息类大数据技术与应用赛项

[时量：300 分钟，试卷号：]

（样卷）

竞赛任务书

场次号：_____ 机位号（工位号、顺序号）：_____ 。

2023 年 12 月 日

注意事项

1. 进入赛位后，您有 15 分钟的准备时间，请认真清点软硬件环境，检查竞赛设备是否完好。

2. 准备时间到，裁判长将发出开赛号令，请接此号令后开始操作。竞赛时间终了前 15 分钟，裁判长将发出时间提醒，竞赛时间到，应立即停机，并终止所有操作。

3. 竞赛连续进行，中途休息、饮食和如厕时间均计算在总时间内，不得中途退场。

4. 您如携带通讯工具及本规程规定的可带工具之外的物品进入竞赛场地，请主动提交裁判。

5. 请严格遵守操作规程，服从裁判指挥，确保设备及人身安全。一旦发生故障，立即停机并立即报告裁判。因人为因素造成的设备故障，裁判长有权决定终止竞赛，如有较严重的违规、违纪、舞弊等现象，裁判组裁定后确定是否取消竞赛成绩；非人为因素造成的设备故障，由裁判长做出裁决。

6. 完成竞赛后，应及时报告裁判，并在裁判指引下，登记确认相关信息，之后不得再进行任何操作。离场时，不得将个人自备物品以外的其他物品带出赛场

一、任务须知

1. 每组参赛队分配容器化竞赛环境、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

二、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（容器环境）、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. 电商实时数据脚本
4. 大数据分析集群环境
5. 数据采集开发环境
6. 实时数据处理开发环境
7. 离线数据处理开发环境

8. 数据可视化开发环境

9. 综合分析报告文档模板

三、具体任务

任务一：大数据平台环境搭建（容器环境）

一、Hadoop 完全分布式安装配置

本环节需要使用 root 用户完成相关配置，安装 Hadoop 需要配置前置环境。命令中要求使用绝对路径，具体要求如下:

1、 将 Master 节点 JDK 安装包解压并移动到/usr/java 路径(若路径不存在,则需新建)，将命令复制并粘贴至对应报告中;

2、 修改/root/profile 文件，设置 JDK 环境变量，配置完毕后在 Master 节点分别执行“java”和“javac”命令，将命令行执行结果分别截图并粘贴至对应报告中;

3、 请完成 host 相关配置，将三个节点分别命名为 master、slave1、slave2，并做免密登录,使用绝对路径从 Master 节点复制 JDK 解压后的安装文件到 Slave1、Slave2 节点，并配置相关环境变量，将全部复制命令复制并粘贴至对应报告中;

4、 在 Master 节点将 Hadoop 解压到/opt 目录下，并将解压包分发至 Slave1、Slave2 节点中,配置好相关环境，初始化 Hadoop 环境 namenode，将初始化命令及初始化结果复制粘贴至对应报告中;

5、 启动 Hadoop 集群,查看 Master 节点 jps 进程,将查看结果复制粘贴至对应报告中。

二、Hive 安装配置

本环节需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体要求如下：

1、 将 Master 节点 Hive 安装包解压到/opt 目录下，将命令复制并粘贴至对应报告中;

2、 设置 Hive 环境变量，并使环境变量生效，并将环境变量配置内容复制并粘贴至对应报告中;

3、 完成相关配置并添加所依赖包，将 MySQL 数据库作为 Hive 元数据库。初始化 Hive 元数据，并通过 schematool 相关命令执行初始化，将初始化结果复制粘贴至对应报告中。

三、Kafka 安装配置

本环节需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体要求如下：

1、 修改 Kafka 的 server.properties 文件，并将修改的内容复制粘贴至对应报告中；

2、 完善其他配置并分发 kafka 文件到 slave1,slave2 中，并在每个节点启动 Kafka，将 Master 节点的 Kafka 启动命令复制粘贴至对应报告中。

任务二：数据采集

一、离线数据采集

编写 Scala 工程代码，将 MySQL 库中表 table1 的数据增量抽取到 Hive 的 ods 库中对应表 table1 中。

1、 抽取库中 table1 的增量数据进入 Hive 的 ods 库中表 table1。根据 ods.table1 表中 modified_time 作为增量字段，只将新增的数据抽入，字段名称、类型不变，同时添加静态分区，分区字段为 etl_date，类型为 String，且值为当前比赛日的前一天日期（分区字段格式为 yyyyMMdd）。使用 hive cli 执行 show partitions ods.table1 命令，将执行结果截图粘贴至对应报告中；

二、实时数据采集

1、 在主节点使用 Flume 采集实时数据生成器 XXXXX 端口的 socket 数据,将数据存入到 Kafka 的 Topic 中，使用 Kafka 自带的消费者消费 Topic 中的数据，查看 Topic 中的前 1 条数据的结果，将查看命令与结果完整的截图粘贴至对应报告中；

2、 实时脚本启动后,在主节点进入到 maxwell 的解压后目录下,配置相关文件并启动，读取主节点 MySQL 数据的 binlog 日志到 Kafka 的 Topic 中。使用 Kafka 自带的消费

者消费 Topic 中的数据，查看 Topic 中的前 1 条数据的结果，将查看命令与结果完整的截图粘贴至对应报告中。

任务三：实时数据处理

一、实时数据清洗

编写 Scala 代码，使用 Flink 消费 Kafka 中 Topic 的数据并进行相应的数据统计计算。

1、 使用 Flink 消费 Kafka 中 topic 的数据，根据数据中不同的表将数据分别分发至 kafka 的 dwd 层的 fact_table1 的 Topic 中，其他的表则无需处理。使用 Kafka 自带的消费者消费 fact_table1（Topic）的前 1 条数据，将结果截图粘贴至对应报告中；

二、实时指标计算

编写 Scala 工程代码，使用 Flink 消费 Kafka 中 dwd 层的 Topic 数据。

1、 使用 Flink 消费 kafka 中的数据，统计商品的 UV 和 PV，将结果写入 HBase 中的表中。使用 Hive cli 查询 HBase 中的表查询出 10 条数据,将结果截图粘贴至对应报告中；

2、 使用 Flink 消费 kafka 中的数据,统计商城每分钟的 GMV,将结果存入 redis 中(value 为字符串格式，仅存 GMV），key 为 store_gmv，使用 redis cli 以 get key 方式获取 store_gmv 值，将每次截图粘贴至对应报告中（每分钟查询一次，查询 3 次）。

任务四：离线数据处理

一、离线数据清洗

编写 Scala 工程代码，将 ods 库中表 table1 抽取到 Hive 的 dwd 库中对应表中。表中有涉及到 timestamp 类型的，均要求按照 yyyy-MM-dd HH:mm:ss，不记录毫秒数，若原数据中只有年月日,则在时分秒的位置添加 00:00:00,添加之后使其符合 yyyy-MM-dd HH:mm:ss。

1、 抽取 ods 库中 table1 表结合 dim_table1 最新分区现有的数据,根据 id 合并数据到 dwd 库中 dim_table1 的分区表（合并是指对 dwd 层数据进行插入或修改，需修改的数据以 id 为合并字段，根据 modified_time 排序取最新的一条），分区字段为 etl_date 且值与 ods 库的相对应表该值相等，同时若 operate_time 为空，则用 create_time 填充,并添加 dwd_insert_user、dwd_insert_time、dwd_modify_user、dwd_modify_time 四列,其中 dwd_insert_user、dwd_modify_user 均填写“user1”。若该条记录第一次进入数仓 dwd 层则 dwd_insert_time、dwd_modify_time 均存当前操作时间，并进行数据类型转换。若该数据在进入 dwd 层时发生了合并修改，则 dwd_insert_time 时间不变,dwd_modify_time 存当前操作时间,其余列存最新的值。使用 hive cli 查询 modified_time 为 XXXX 年 XX 月 X 日当天的数据，将结果截图粘贴至对应报告中；

二、离线指标计算

1、 编写 Scala 工程代码，根据 dwd 的订单表，求各省份下单时间为 XXXX 年的支付转化率，并将计算结果写入 clickhouse 的 ds_result 库的表。在 Linux 的 clickhouse 命令行中根据 ranking 字段查询出转化率前三的省份，将 SQL 语句与执行结果截图粘贴至对应报告中；

任务五：数据可视化

1、 用柱状图展示消费额最高的省份

2、 用饼状图展示各地区消费能力

3、 用折线图展示每年上架商品数量的变化

任务六：综合分析报告

根据项目要求，完成综合分析报告编写。

四、竞赛结果提交要求

1、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。

2、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）。