**2023年浙江省职业院校技能大赛**

**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）、不超过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。

3、竞赛时间6小时。

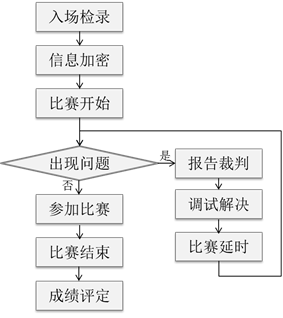
五、竞赛流程

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。

1. **竞赛时间安排**

|  |  |  |
| --- | --- | --- |
| **日期** | **时间** | **内容** |
| 4月  27日 | 13:00-15:00 | 各参赛队报到 |
| 15:00-16:00 | 领队会、赛前说明 |
| 15:30-16:30 | 选手熟悉赛场 |
| 4月  28日 | 7:40-8:40 | 赛场检录，抽签加密 |
| 8:40-9:00 | 竞赛选手进入赛位，检查设备，现场裁判发放任务书、宣布竞赛注意事项 |
| 9:00-15:00 | 竞赛选手完成竞赛任务 |
| 15:00-18:00 | 对选手提交的结果文件进行评分 |

注：最终竞赛流程以参赛指南为准。

1. **竞赛流程**

六、竞赛命题

技能大赛的命题工作由赛项执委会指定的命题专家组负责。

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. 参赛选手须为2023年度同校在籍高职学生组成，性别和年级不限。凡在往届全国职业院校技能大赛中获一等奖的选手，不能再参加同一项目同一组别的比赛。参赛选手由省教育厅、省人力资源社会保障厅负责。

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参赛证件进行检录。

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. 竞赛成绩经复核无误后，经裁判长、监督人员审核签字后，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。

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、指导老师休息区。

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. 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 |

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 | 每支参赛队伍一套 | 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**

采用四合天地大数据实训管理系统（北京四合天地科技有限公司研发）。

**（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 |

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”，设裁判长1名,全面负责竞赛裁判和管理工作。

裁判员根据竞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。

**（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 |

**（五）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。

**（六）成绩公布方法**

竞赛成绩经复核无误后，经裁判长、监督人员审核签字后，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年浙江省职业院校技能大赛（高职组）“大数据技术与应用”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：大数据技术与应用赛项竞赛试题（样卷）

# 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、饮水饮食、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 |
| 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。

# 任务须知

1. 每组参赛队分配一台服务器，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。

# 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。

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）。