**2023-2024学年河北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

“十四五”时期，大数据产业对经济社会高质量发展的赋能作用更加突显，大数据已成为催生新业态、激发新模式、促进新发展的技术引擎。习近平总书记指出“大数据是信息化发展的新阶段”，“加快数字化发展，建设数字中国”成为《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》的重要篇章。

本赛项旨在落实国家“建设数字中国”战略，协同推动大数据相关产业的创新与发展，大力推进大数据技术及相关专业的技术技能型人才培养，全面提升相关专业毕业生的综合能力，展现选手团队合作、工匠精神等职业素养，赋能经济社会高质量发展。竞赛内容结合当前大数据相关产业中的新技术、新要求如数据湖、OLAP数据库应用等，全面检验参赛选手的工程实践能力和创新能力，推进教学过程与生产过程对接、课程内容与职业标准对接、专业设置与产业需求对接，促进职普融通、产教融合、科教融汇，引领专业建设和教学改革。竞赛内容围绕大数据相关产业岗位的实际技能要求进行设计，通过竞赛搭建校企合作的平台，强化竞赛成果转化，促进相关教材、资源、师资、认证、实习就业等方面的全方位建设，满足产教协同育人目标，为国家战略规划提供大数据领域高素质技能型人才。

三、竞赛内容

本赛项涉及的典型工作任务包括大数据平台搭建（容器环境）、离线数据处理、数据挖掘、数据采集与实时计算、数据可视化、综合分析、职业素养，引入行业内较为前沿的数据湖架构作为创新、创意的范围与方向，考查的技术技能如下：

（一）大数据平台搭建（容器环境）：Docker容器基础操作、Hadoop完全分布式安装配置、Hadoop HA安装配置、Spark on Yarn安装配置、Flink on Yarn安装配置、Hive安装配置、Flume安装配 置、ZooKeeper安装配置、Kafka安装配置、HBase分布式安装配置、ClickHouse单节点安装配置、Hudi安装配置。

（二）离线数据处理：Scala应用开发、Pom文件配置、Maven本地仓库配置使用、基于Spark的数据清洗处理方法、基于Hive的数据清洗处理方法、基于Hudi的数据清洗处理方法、数据仓库基本架构及概念、数据湖基本架构及概念、MySQL基本操作、ClickHouse基本操作、Azkaban基本操作、DolphinScheduler基本操作。

（三）数据挖掘：特征工程应用、Spark ML机器学习库应用开发、推荐算法的召回和排序、回归模型、聚类模型、决策树模型、随机森林模型应用。

（四）数据采集与实时计算：Scala应用开发、Pom文件配置、Maven本地仓库配置使用、基于Flume及Kafka的数据采集方法、基于Flink的实时数据处理方法、HBase基本操作、Redis基本操作、MySQL基本操作。

（五）数据可视化：Vue.js框架应用开发、ECharts组件应用开发，会使用 ECharts绘制柱状图、折线图、折柱混合图、玫瑰图、气泡图、饼状图、条形图、雷达图、散点图等图表。

（六）综合分析：依据整体项目过程，在综合理解业务的基础上，根据题目要求进行综合分析。

（七）职业素养：团队分工明确合理、操作规范、文明竞赛。

1、竞赛内容结构、成绩比例如下：

表 3-1 竞赛内容结构和成绩比例

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 竞赛任务 | 成绩比例 | 考核内容 |
| 1 | 大数据平台搭建（容器环境） | 15% | 选手在容器环境下对大数据平台及相关组件的安装、配置、可用性验证等内容。 |
| 2 | 离线数据处理 | 25% | 选手对Hadoop 平台、Spark平台、Hive数据仓库、Hudi 数据湖、任务调度工具等的综合应用能力，使用Scala开发语言，完成离线数据抽取、数据清洗、数据指标统计等操作，并存入MySQL、ClickHouse中。 |
| 3 | 数据挖掘 | 10% | 选手运用常用的机器学习方法对数据进行数据挖掘分析。 |
| 4 | 数据采集与实时计算 | 20% | 选手对 Flink 平台、Flume 组件、Kafka组件等的综合应用能力，基于Flume和Kafka进行实时数据采集，使用Scala开发语言，完成实时数据流相关数据指标的分析、计算等操作，并存入HBase、Redis、MySQL中。 |
| 5 | 数据可视化 | 15% | 选手基于前端框架Vue.js和后端REST风格 的数据接口，使用JavaScript语言将数据分析结果以图表的形式进行呈现、统计。 |
| 6 | 综合分析 | 10% | 选手对大数据技术的业务分析、技术分析及报告撰写能力。 |
| 7 | 职业素养 | 5% | 团队分工明确合理、操作规范、文明竞赛。 |

四、竞赛方式

本竞赛为线下比赛，组队方式为师生同赛，具体要求如下：

（一）参赛学生须为高等职业学校专科、高等职业学校本科全日制在籍学生，五年制高职四、五年级学生也可报名参赛；参赛教师须为校内专任教师，并提供近半年的社保或纳税证明。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。

（二）每支参赛队由4名选手组成，其中1名教师，3名学生。本赛项为师生同赛不设指导教师，报名获得确认后不得随意更换。

（三）本赛项为单一场次，所有参赛队在现场根据给定的任务说明，在8小时内相互配合，采用小组合作的形式完成任务，最后以提交的结果文档作为最终评分依据。

五、竞赛流程

根据竞赛任务要求，参赛队伍在8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**

详见后续通知

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.学生须为高等职业学校专科、高等职业学校本科、技师学院全日制在籍学生，五年制高职四、五年级学生也可报名参赛。凡在往届全国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。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.赛项成绩解密后，在指定地点，以纸质形式向全体参赛队进行公布，成绩无异议后，在闭赛式上予以宣布。

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、服务区。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**

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **设备名称** | **数量** | **备注** |
| 1 | 服务器 | 每组1台 | CPU：Intel 至强银牌 4210 及以上内存：不少于 128GB硬盘：不少于 1TB网卡：千兆 |
| 2 | PC 机 | 每组4台 | CPU：i5 及以上内存：不少于 16GB硬盘：不少于NVME 500GB或SATA 1T显示器：1920\*1080 及以上 |
| 3 | 交换机 | 每组1台 | 8口及以上千兆交换机 |

**（二）软件平台**

由2023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。

系统基于自研Docker容器编排管理引擎，运用云原生和容器技术构建训练环境，支持快速创建训练环境。实现每个学生环境互相隔离、训练过程互不干扰，教师可以一键操作即可创建一套全新的环境供学生进行备赛训练，方便学生高效的完成训练操作的同时，大幅降低了教师组织训练的难度和成本。具体功能如下：

1.本系统运行在开源操作系统Linux平台下，应使用浏览器/服务器模式提供服务,用户使用最新版本的谷歌浏览器访问系统。

2．系统支持包括管理员、教师、学生三种角色。管理员负责系统配置维护、镜像环境维护等工作；教师负责小组维护、训练任务维护等工作；学生参与并完成训练。

管理员功能：

3．专业管理：管理员可自主创建专业，编辑专业相关内容，如：新增（编辑）专业，删除专业。

4．班级管理：管理员可自主创建班级，编辑班级相关内容，如：新增（编辑）班级，删除班级。

5．用户管理：管理员可自主创建用户，编辑用户相关内容，如：新增（编辑）用户的账号、姓名、手机号、角色、状态、密码等，同时也可对用户进行删除操作。

6．镜像环境：管理员可维护镜像环境，可通过网页上传Dockerfile文件来完成镜像环境的创建，也可从本地镜像仓库进行镜像同步，并可对镜像环境执行编辑等操作。

7．环境配置：管理员依据镜像环境的具体应用场景，通过添加环境配置来完善该镜像所需的CPU、内存、磁盘空间等配置信息，并可对环境配置信息进行编辑和删除。

8．服务器配置：管理员可配置计算节点服务器相关信息，包括新增服务器名称、服务器IP地址、CPU核数、内存大小、磁盘空间，也可对服务器配置信息进行编辑和删除。

9．训练资源监控：管理员可监控正在进行的训练资源，监控正在进行的训练状态、使用人数、占用资源等信息，并可对相关的环境执行启动、挂起等操作。

10．角色管理：管理员可自主创建角色，编辑角色相关内容，如：添加（编辑）角色名称、角色备注（描述）、状态、权限分配（权限分配依据角色名称分配相应权限）等，同时也可对角色进行删除操作。

11．日志管理：管理员可查看当前系统操作日志和登录日志。操作日志主要记录操作的账户、操作模块、时间、IP等信息。登录日志主要记录登录的用户、IP地址、登录状态等信息。

12．系统设置：可自定义系统名称，如浏览器标题、浏览器LOGO、登录页标题、首页标题、首页LOGO等。

教师功能：

13．训练模块：教师可自主创建训练模块，编辑相应的训练内容。内容包括训练的名称、介绍、使用状态、该训练指定的一个或多个实训环境、可供下载的资源包、任务设置的具体形式包括上传任务说明、任务参考等。

14．小组管理：教师可自主创建小组，并对小组进行编辑、启用、禁用等操作。已启用的小组可以进行添加成员、设置任务参考操作。在编辑具体小组时，通过选择该小组所包含的成员，内容上选择所用到的训练模块，完成小组的整体设置。

15．训练监控：教师可查看各小组的训练情况，选择启动、挂起、继续或结束训练环境。通过“进入”按钮，可进入小组的某个环境进行指导操作，也可对小组的某个训练环境执行启动、停止、重置或重启。

16．训练报告：对已启用小组，记录每个小组中，学生具体的训练情况，查阅学生提交上传的报告、附件等信息，编辑每个学生对应的训练成绩。

17．个人实验环境：教师可在个人实验环境中选择镜像环境进行个人实验或练习等操作。

18．消息通知：当学生完成训练报告的提交或附件上传时，教师会收到消息通知并可通过消息通知直接找到训练报告或附件。

19．个人资料：教师可修改个人头像及密码。

学生功能：

20．学生登录平台，可进入对应的训练模块，在线查看任务说明、任务参考，并进入训练环境进行操作，可在浏览器中访问自己的训练环境，也可通过SSH工具直连相应训练环境容器的ip地址进入该环境，支持通过环境的ip地址访问环境中部署的相应服务，也可对环境进行重置、重启、上传、下载等操作，同时学生可以下载相关的资源文件，提交个人的训练报告，上传相关附件完成训练。

21．个人中心：学生可修改个人头像及密码，也可查看自己的训练记录。

**（三）软件环境**

|  |  |  |
| --- | --- | --- |
| **设备类型** | **软件类别** | **软件名称、版本号** |
| 服务器 | 大数据集群操作系统 | CentOS 7 |
| 容器环境 | Docker-CE 20.10 |
| 大数据平台组件 | Hadoop 3.1.3 |
| Yarn 3.1.3 |
| ZooKeeper 3.5.7 |
| Hive 3.1.2 |
| Hudi 0.12.0 |
| ClickHouse 21.9.4 |
| JDK 1.8 |
| Flume 1.9.0 |
| Kafka 2.4.1 |
| Spark 3.1.1 |
| Flink 1.14.0 |
| Redis 6.2.6 |
| HBase 2.2.3 |
| Azkaban 3.84.4 |
| DolphinScheduler 3.1.4 |
| 关系型数据库 | MySQL 5.7 |
| PC 机 | PC 操作系统 | Ubuntu18.04 64 位 |
| 浏览器 | Chrome |
| 开发语言 | Scala 2.12 |
| JavaScript |
| 开发工具 | IDEA 2022（Community Edition） |
| Visual Studio Code 1.69 |
| SSH 工具 | Asbru-cm 或 Ubuntu SSH 客户端 |
| 数据库工具 | MySQL Workbench |
| 接口测试工具 | Postman |
| 数据可视化框架及组件 | Vue.js 3.2 |
| ECharts 5.1 |
| 截图工具 | Ubuntu 系统自带 |
| 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”。

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 | 子任务 | 主要知识与技能点 | 分值 |
| 任务 A:大数 据平台搭建（容器环境） | 子任务一：Hadoop完全分布式安装配置 | Hadoop 完全分布式下的 JDK 的解压安装、JDK 环境变量配置、节点配置、Hadoop 配置文件修改、运行测试等 | 7 |
| 子任务二：Spark on Yarn 安装配置 | Spark的解压安装、环境变量配置、on Yarn 配置、运行测试等 | 4 |
| 子任务三：HBase分布式安装配置 | HBase的解压安装、环境变量配置、运行测试等 | 4 |
| 小计 | 15 |
| 任务 B:离线数据处理 | 子任务一：数据抽取 | 从 MySQL中进行离线数据抽取到 Hive、Hudi的相关操作 | 6 |
| 子任务二：数据清洗 | 从 ods 到 dwd的数据清洗，包括全量数据 取、数据合并、数据排序、去重、数据类型转换等操作 | 6 |
| 子任务三：指标计算 | 在 dwd、dws层进行任务调度，对数据进行相关数据指标的统计、计算等操作,将结果存入 MySQL、ClickHouse中 | 13 |
| 小计 | 25 |
| 任务 C:数据挖掘 | 子任务一：特征工程 | 对推荐系统的数据集进行特征提取及数据预处理等操作 | 5 |
| 子任务二：推荐系统 | 基于用户的推荐系统设计开发操作 | 5 |
| 小计 | 10 |
| 任务 D:数据采集与实时计算 | 子任务一：实时数据采集 | 基于Flume和 Kafka的实时数据采集，包括 Flume 采集配置、数据注入 Kafka等操作 | 8 |
| 子任务二：使用 Flink处理 Kafka 中的数据 | 使用Flink消费Kafka中的数据进行实时 计算，包括Kafka基本操作、实时数据统计计算、HBase 基本操作、Redis 基本操作、MySQL基本操作等。 | 12 |
| 小计 | 20 |
| 任务 E：数据可视化 | 子任务一：用柱状图展示消费额最高的省份 | 正确使用 Vue.js 框架，结合 ECharts绘制柱状图。 | 3 |
| 子任务二：用柱状图展示消费额最低的省份 | 正确使用 Vue.js框架，结合ECharts绘制柱状图。 | 3 |
| 子任务三：用折线图展示每年上架商品数量变化 | 正确使用 Vue.js 框架，结合 ECharts绘制折线图。 | 2 |
|  |
| 子任务四：用条形图展示平均消费额最高的省份 | 正确使用 Vue.js框架，结合ECharts绘制条形图。 | 3 |
| 子任务五：用折柱混合图展示省份平均消费额和地区平均消费额 | 正确使用 Vue.js 框架，结合 ECharts 绘制折柱混合图。 | 4 |
| 小计 | 15 |
| 任务 F:综合分析 | 子任务一：Flink 有哪些重启策略？各个重启策略如何配置？ | 正确分析 Flink 的重启策略 | 4 |
| 子任务二：Hadoop 有哪些类型的调度器？简要说明其工作方法。 | 正确分析 Hadoop 的调度器 | 3 |
| 子任务三：分析下一年度的建仓目的地。 | 根据可视化图表，合理分析下一年度的建仓目的地 | 3 |
| 小计 | 10 |
| 任务 G:职业素养 | 考察职业素养 | 竞赛团队分工明确合理、操作规范、文明竞赛 | 5 |
| 小计 | 5 |
| 总分 | 100 |

**（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、赛场预案

赛场备用工位：赛场提供占总参赛队伍10%的备用工位。

竞赛系统可靠性：竞赛系统使用的服务器应进行冗余，数据库、存储应使用高可用架构。提前开始运行，经过多次压力测试，由学校组织的真实竞赛环境测试。

竞赛备用服务器：现场提供占总参赛队伍10%的备用服务器。

现场应急预案详情，如下：

**（一）服务器问题预案**

若服务器在比赛过程中出现卡顿、死机等情况，参赛选手举手示意裁判，在裁判与技术支持人员确定情况后，可更换服务器。更换服务器的等待时间，可在比赛结束后延时。

**（二）交换机问题预案**

若交换机在比赛过程中出现传输速度慢或无故中断等情况，参赛选手举手示意裁判，在裁判与技术支持人员确定情况后，可更换交换机。更换交换机的等待时间，可在比赛结束后延时。

**（三）PC机问题预案**

若PC机在比赛过程中出现死机、蓝屏等现象（重启后无法解决），参赛选手举手示意裁判，在裁判与技术支持人员确定情况后，可更换备用工位或更换PC机进行答题。

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**

1.2024年河北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。

十四、竞赛观摩

**（一）视频观摩**

赛场外设置开放式观摩区，向媒体、企业代表、院校师生等社会公众开放，通过室外大屏幕对赛场进行直播，同时还可以通过竞赛系统进度监控图实时观看选手答题进度。

**（二）组织安排**

在竞赛开始1小时之后，由承办校组织并派人带领媒体、专家、企业代表、院校师生等进入赛场外的开放式观摩区，按照指定路线进行观摩。

**（三）纪律要求**

为保证大赛顺利进行，在观摩期间应遵循以下纪律要求：

1.除与竞赛直接有关工作人员、裁判员、参赛选手外，其余人员均为观摩观众。

2.不得违反职业院校技能大赛规定的各项纪律。

3.观摩人员需批准，佩戴观摩证件，遵循观摩区的工作人员指挥。

4.文明观摩，保持观摩区清洁，不得大声喧哗，杜绝各种违反观摩秩序的不文明行为。

十五、竞赛直播

**（一）直播方式**

1.赛场内部署无盲点录像设备，能实时录制并播送赛场情况。

2.赛场外有大屏幕或投影，同步显示赛场内竞赛状况。

3.赛场外通过大屏幕或投影，实时展示竞赛系统选手答题进度。

**（二）直播安排**

1.对赛项赛场准备、开赛式和闭赛式、比赛期间进行录像。

2.从选手进入赛场开始，全程进行赛场实时录像直播。

3.从比赛开始时到比赛结束，全程进行竞赛系统的进度监控直播。

**十六、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4名参赛选手组成，须为同校在籍学生和教师，同一学校的报名参赛队伍不超过2支；

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，经赛项执委会办公室核实后予以更换。允许队员缺席比赛。

5.参赛队按照赛项竞赛规程安排，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，须为参赛选手购买竞赛期间的人身意外伤害保险。

**（二）领队须知**

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.领队务必带好有效身份证件，在活动过程中佩戴“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，由领队负责以书面形式向赛项仲裁委员会反映，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）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。

**附件一、竞赛样题**

背景描述

大数据时代背景下，电商经营模式发生很大改变。在传统运营模式中，缺乏数据积累，人们在做出一些决策行为过程中，更多是凭借个人经验和直觉，发展路径比较自我封闭。而大数据时代，为人们提供一种全新的思路，通过大量的数据分析得出的结果将更加现实和准 确。商家可以对客户的消费行为信息数据进行收集和整理，比如消费 者购买产品的花费、选择产品的渠道、偏好产品的类型、产品回购周期、购买产品的目的、消费者家庭背景、工作和生活环境、个人消费观和价值观等。通过数据追踪，知道顾客从哪儿来，是看了某网站投放的广告还是通过朋友推荐链接，是新访客还是老用户，喜欢浏览什么产品，购物车有无商品，是否清空，还有每一笔交易记录，精准锁定一定年龄、收入、对产品有兴趣的顾客，对顾客进行分组、标签化，通过不同标签组合运用，获得不同目标群体，以此开展精准推送。

因数据驱动的零售新时代已经到来，没有大数据，我们无法为消费者提供这些体验，为完成电商的大数据分析工作，你所在的小组将应用大数据技术，以 Scala 作为整个项目的基础开发语言，基于大数据平台综合利用 Hive、Spark、Flink、Vue.js 等技术，对数据进行处理、分析及可视化呈现，你们作为该小组的技术人员，请按照下面任务完成本次工作。

任务A：大数据平台搭建(容器环境)（15分)

环境说明：

|  |
| --- |
| 服务端登录地址详见各任务服务端说明。补充说明：宿主机及各容器节点可通过 Asbru 工具或 SSH 客户端进行SSH访问。 |

子任务一：Hadoop 完全分布式安装配置

本任务需要使用 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. 从宿主机/opt目录下将文件 hadoop-3.1.3.tar.gz、jdk-8u212- linux-x64.tar.gz 复制到容器Master 中的/opt/software 路径中 （若路径不存在，则需新建），将 Master 节点 JDK 安装包解压到 /opt/module 路径中(若路径不存在，则需新建)，将 JDK 解压命 令复制并粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下；
2. 修改容器中/etc/profile 文件，设置 JDK 环境变量并使其生效， 配置完毕后在 Master 节点分别执行 “ java -version”和 “ java c ”命令，将命令行执行结果分别截图并粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下；
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，用 scp 命令并使用绝对路径从 Master 复 制 JDK 解压后的安装文件到 slave1、slave2 节点（若路径不存在， 则需新建），并配置 slave1、slave2 相关环境变量，将全部 scp 复制 JDK 的命令复制并粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下；
4. 在容器 Master 将 Hadoop 解压到/opt/module (若路径不存在，则需新建)目录下，并将解压包分发至 slave1、slave2 中，其中 master、slave1、slave2 节点均作为 datanode，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截图 （截取初始化结果日志最后 20 行即可）粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下；
5. 启动 Hadoop 集群（包括 hdfs 和 yarn），使用 jps 命令查看 Mas ter 节点与 slave1 节点的 Java 进程，将 jps 命令与结果截图粘 贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下。

子任务二：Spark on Yarn 安装配置

本任务需要使用 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. 从宿主机/opt目录下将文件 spark-3.1.1-bin-hadoop3.2.tgz 复制到容器 Master中的/opt/software（若路径不存在，则需新建）中，将 Spark 包解压到/opt/module 路径中(若路径不存在，则需新建)，将完整解压命令复制粘贴至客户端桌面【Release\任务 A提交结果.docx】中对应的任务序号下；
2. 修改容器中/etc/profile文件，设置 Spark 环境变量并使环境变 量生效，在/opt 目录下运行命令 spark-submit --version，将命令与结果截图粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下；
3. 完成 on yarn 相关配置，使用 spark on yarn 的模式提交$SPARK\_HOME/examples/jars/spark-examples\_2.12-3.1.1.jar 运行的 主类为 org.apache.spark.examples.SparkPi，将运行结果截图 粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下（截取 Pi 结果的前后各 5 行）。

（运行命令为：spark-submit --master yarn --class org.ap ache.spark.examples.SparkPi $SPARK\_HOME/examples/jars/spark-example\_2.12-3.1.1.jar）

子任务三：HBase 分布式安装配置

本任务需要使用 root用户完成相关配置，安装HBase 需要配置 Hadoop 和 ZooKeeper 等前置环境。命令中要求使用绝对路径，具体要求如下:

1. 从宿主机/opt 目录下将文件 apache-zookeeper-3.5.7-bin.tar.gz、hbase-2.2.3-bin.tar.gz 复制到容器 Master 中的/opt/software 路径中 (若路径不存在，则需新建) ，将 zookeeper、hbase 安装包解压到/opt/module 目录下，将 HBase 的解压命令复制并 粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下；
2. 完成 ZooKeeper 相关部署，用 scp 命令并使用绝对路径从容器ma ster 复制 HBase 解压后的包分发至 slave1、slave2 中，并修改相关配置，配置好环境变量，在容器Master 节点中运行命令hbase version，将全部复制命令复制并将 hbase version 命令的结果截图粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对 应的任务序号下;
3. 启动 HBase 后在三个节点分别使用 jps 命令查看，并将结果分别 截图粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应 的任务序号下；正常启动后在 hbase shell 中查看命名空间，将 查看命名空间的结果截图粘贴至客户端桌面【Release\任务 A 提交结果.docx】中对应的任务序号下。

任务B：离线数据处理(25分)

环境说明：

|  |
| --- |
| 服务端登录地址详见各任务服务端说明。补充说明：各节点可通过 Asbru 工具或 SSH 客户端进行 SSH 访问；主节点 MySQL 数据库用户名/密码：root/123456（已配置远程连接）；Hive 的配置文件位于/opt/apache-hive-2.3.4-bin/conf/Spark 任务在 Yarn 上用 Client 运行，方便观察日志。 |

子任务一：数据抽取

编写 Scala 代码，使用 Spark 将 MySQL 的 shtd\_store 库中表 user\_info、sku\_info、base\_province、base\_region、order\_info、o r der\_detail 的数据增量抽取到 Hive 的 ods 库中对应表 user\_info、 sku\_info、base\_province、base\_region、order\_info、order\_detail 中。

1. 抽取 shtd\_store 库中 user\_info 的增量数据进入 Hive 的 ods 库中表 user\_info。根据 ods.user\_info 表中 operate\_time 或 create\_time 作为增量字段(即 MySQL 中每条数据取这两个时间中较 大的那个时间作为增量字段去和 ods 里的这两个字段中较大的时 间进行比较)，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.user\_info 命令，将结果截图 粘贴至客户端桌面【Release\任务B提交结果.docx】中对应的任务序号下；
2. 抽取 shtd\_store 库中 sku\_info 的增量数据进入 Hive 的 ods 库中表 sku\_info。根据 od s.sku\_info 表中 create\_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 partition s ods.sku\_info 命令，将结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；
3. 抽取 shtd\_store 库中 base\_province 的增量数据进入 Hive 的 od s 库中表 base\_province。根据 ods.base\_province 表中 id 作为 增量字段，只将新增的数据抽入，字段名称、类型不变并添加字 段 create\_time 取当前时间，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.base\_province 命令，将结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；
4. 抽取 shtd\_store 库中 base\_region 的增量数据进入 Hive 的 ods 库中表 base\_region。根据 ods.base\_region 表中 id 作为增量字段，只将新增的数据抽入，字段名称、类型不变并添加字段 crea te\_time 取当前时间，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.base\_region 命令，将结果截图粘贴至客户端桌面【Release\任务 B提交结果.docx】中对应的任务序号下；
5. 抽取 shtd\_store 库中 order\_info 的增量数据进入 Hive 的 ods 库中表 order\_info，根据 ods.order\_info 表中 operate\_time 或 create\_time 作为增量字段(即 MySQL 中每条数据取这两个时间中 较大的那个时间作为增量字段去和 ods 里的这两个字段中较大的时间进行比较)，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.order\_info 命令，将结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；
6. 抽取 shtd\_store 库中 order\_detail 的增量数据进入 Hive 的 ods库中表 order\_detail，根据 ods.order\_detail 表中 create\_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.order\_detail 命令，将结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；

子任务二：数据清洗

编写 Scala 代码，使用 Spark 将 ods 库中相应表数据全量抽取到 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 user\_info 表中昨天的分区（子任务一生成的分区） 数据，并结合 dim\_user\_info 最新分区现有的数据，根据 id 合并 数据到 dwd 库中 dim\_user\_info 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 operate\_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 d wd 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 执行 show partitions dwd.dim\_user\_info 命令，将结果截图粘贴至客户端桌面【Re lease\任务 B 提交结果.docx】中对应的任务序号下；
2. 抽取 ods 库 sku\_info 表中昨天的分区（子任务一生成的分区）数 据，并结合 dim\_sku\_info 最新分区现有的数据，根据 id 合并数据到 dwd 库中 dim\_sku\_info 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 create\_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，并添加 dwd\_insert\_user、dwd\_insert\_time、 dwd\_modify\_user、dwd\_modify\_time 四列,其中 dwd\_insert\_use r、dwd\_modify\_user 均填写 “user1 ”。若该条数据第一次进入 数仓 dwd 层则 dwd\_insert\_time、dwd\_modify\_time 均填写当前操 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 查询表 dim\_s ku\_info 的字段 id、sku\_desc、dwd\_insert\_user、dwd\_modify\_ time、etl\_date，条件为最新分区的数据，id 大于等于 15 且小于等于 20，并且按照 id 升序排序，将结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；
3. 抽取 ods 库 base\_province 表中昨天的分区（子任务一生成的分区）数据，并结合 dim\_province 最新分区现有的数据，根据 id 合并数据到 dwd 库中 dim\_province 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 create\_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，并添加 dwd\_insert\_user、dwd\_in ser t\_time、dwd\_modify\_user、dwd\_modify\_time 四列,其中 dwd\_insert\_user、dwd\_modify\_user 均填写 “user1 ”。若该条数据第 一次进入数仓 dwd 层则 dwd\_insert\_time、dwd\_modify\_time 均填写当前操作时间，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 在表 dwd.dim\_province 最新分区中，查询该分区中数据的条数，将结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；
4. 抽取 ods 库 base\_region 表中昨天的分区（子任务一生成的分区） 数据，并结合 dim\_region 最新分区现有的数据，根据 id 合并数 据到 dwd 库中 dim\_region 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 create\_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，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”。若该条数据第一次进入数仓 dwd 层则 dwd\_insert\_time、dwd\_modify\_time 均填写当前操作时 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 在表 dwd.dim\_reg ion 最新分区中，查询该分区中数据的条数，将结果截图粘贴至 客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；
5. 将 ods 库中 order\_info 表昨天的分区（子任务一生成的分区）数据抽取到 dwd 库中 fact\_order\_info 的动态分区表，分区字段为 etl\_date，类型为 String，取 create\_time 值并将格式转换为 yyyyMMdd，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\_us er 均填写 “user1”，dwd\_insert\_time、dwd\_modify\_time 均填写当前操作时间，并进行数据类型转换。使用 hive cli 执行 show partitions dwd.fact\_order\_info 命令，将结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；
6. 将 ods库中 order\_detail表昨天的分区（子任务一中生成的分区） 数据抽取到 dwd 库中 fact\_order\_detail 的动态分区表，分区字段为 etl\_date，类型为 String，取 create\_time 值并将格式转换为 yyyyMMdd，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”，dwd\_insert\_time、dwd\_mod ify\_time 均填写当前操作时间，并进行数据类型转换。使用 hiv e cli 执行 show partitions dwd.fact\_order\_detail 命令，将 结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下。

子任务三：指标计算

编写 Scala 代码，使用 Spark 计算相关指标。

1.本任务基于以下 2、3、4 小题完成，使用 Azkaban 完成第 2、3、4 题任务代码的调度。工作流要求，使用 shell 输出 “开始 ”作为工作流的第一个 job（job1），2、3、4 题任务为串行任务且它 们依赖 job1 的完成（命名为 job2、job3、job4），job2、job3、job4 完成之后使用 shell 输出 “结束”作为工作流的最后一个 job（endjob），endjob 依赖 job2、job3、job4，并将最终任务调 度完成后的工作流截图，将截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| provinceid | in t | 省份表主键 |  |
| provincename | text | 省份名称 |  |
| region id | in t | 地区表主键 |  |
| regionname | text | 地区名称 |  |
| totalconsumption | double | 订单总金额 | 当月订单总金额 |
| totalorder | in t | 订单总数 | 当月订单总数 |
| year | in t | 年 | 订单产生的年 |
| month | in t | 月 | 订单产生的月 |

2.根据 dwd 层表统计每个省份、每个地区、每个月下单的数量和下 单的总金额，存入 MySQL 数据库 shtd\_result 的 provinceeverymon th 表中（表结构如下），然后在 Linux 的 MySQL 命令行中根据 订单总数、订单总金额、省份表主键均为降序排序，查询出前 5 条，将 SQL 语句复制粘贴至客户端桌面【Release\任务 B 提交结 果.docx】中对应的任务序号下，将执行结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下；

3.请根据 dwd 层表计算出 2020 年 4 月每个省份的平均订单金额和所有省份平均订单金额相比较结果（ “高/低/相同”）,存入 MySQL 数据库 shtd\_result 的 provinceavgcmp 表（表结构如下）中，然后在 Linux 的 MySQL 命令行中根据省份表主键、该省平均订单 金额均为降序排序，查询出前 5 条，将 SQL 语句复制粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下，将执行结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下;

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| province id | in t | 省份表主键 |  |
| provincename | text | 省份名称 |  |
| provinceavgconsumption | double | 该省平均订单金额 |  |
| allprovinceavgconsumption | double | 所有省平均订单金额 |  |
| comparison | text | 比较结果 | 该省平均订单金额和所有省平均订单金额比较结果，值为：高/低/相同 |

4.根据 dwd 层表统计在两天内连续下单并且下单金额保持增长的用户，存入 MySQL 数据库 shtd\_result 的 usercontinueorder 表(表结构如下)中，然后在 Linux 的 MySQL 命令行中根据订单总数、订单总金额、客户主键均为降序排序，查询出前 5 条，将 SQL 语句 复制粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应 的任务序号下，将执行结果截图粘贴至客户端桌面【Release\任务 B 提交结果.docx】中对应的任务序号下。

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| userid | in t | 客户主键 |  |
| username | text | 客户名称 |  |
| day | text | 日 | 记录下单日的时间，格式为yyyyMMdd\_yyyyMMdd例如： 20220101\_20220102 |
| totalconsumption | double | 订单总金额 | 连续两天的订单总金额 |
| totalorder | in t | 订单总数 | 连续两天的订单总数 |

任务 C：数据挖掘（10 分）

环境说明：

|  |
| --- |
| 服务端登录地址详见各任务服务端说明。补充说明：各节点可通过 Asbru 工具或 SSH 客户端进行 SSH 访问； 主节点 MySQL 数据库用户名/密码：root/123456（已配置远程连接）；Hive 的配置文件位于/opt/apache-hive-2.3.4-bin/conf/Spark 任务在 Yarn 上用 Client 运行，方便观察日志。该任务均使用 Scala 编写，利用 Spark 相关库完成。 |

子任务一：特征工程

1.根据 Hive 的 dwd 库中相关表或 MySQL 中 shtd\_store 中相关表（o r der\_detail、sku\_info），计算出与用户 id 为 6708 的用户所购 买相同商品种类最多的前 10 位用户（只考虑他俩购买过多少种相 同的商品，不考虑相同的商品买了多少次），将 10 位用户 id 进 行输出，若与多个用户购买的商品种类相同，则输出结果按照用户 id 升序排序，输出格式如下，将结果截图粘贴至客户端桌面【Release\任务 C 提交结果.docx】中对应的任务序号下；结果格式如下：

-----------------相同种类前 10 的 id 结果展示为：------------------

1,2,901,4,5,21,32,91,14,52

2.根据 Hive 的 dwd 库中相关表或 MySQL中 shtd\_store 中相关商品表（sku\_info），获取 id、spu\_id、price、weight、tm\_id、category3\_id 这六个字段并进行数据预处理，对 price、weight 进 行规范化(StandardScaler)处理，对 spu\_id、tm\_id、category3\_id 进行 one-hot 编码处理（若该商品属于该品牌则置为 1，否则置为 0）,并按照 id 进行升序排序，在集群中输出第一条数据前 10 列（无需展示字段名），将结果截图粘贴至客户端桌面【Re lease\任务 C 提交结果.docx】中对应的任务序号下。

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 中文含义 | 备注 |
| id | double | 主键 |  |
| price | double | 价格 |  |
| weight | double | 重量 |  |
| spu\_id#1 | double | spu\_id 1 | 若属于该 spu\_id，则内容为 1 否则为 0 |
| spu\_id#2 | double | spu\_id 2 | 若属于该 spu\_id，则内容为 1 否则为 0 |
| ..... | double |  |  |
| tm\_id#1 | double | 品牌 1 | 若属于该品牌，则内容为 1 否则为 0 |
| tm\_id#2 | double | 品牌 2 | 若属于该品牌，则内容为 1 否则为 0 |
| …… | double |  |  |
| category3\_id#1 | double | 分类级别 31 | 若属于该分类级别 3，则内容为 1 否则为 0 |
| category3\_id#2 | double | 分类级别 32 | 若属于该分类级别 3，则内容为 1 否则为 0 |
| …… |  |  |  |

结果格式如下：

------------第一条数据前 10 列结果展示为：----------

1.0,0.892346,1.72568,0.0,0.0,0.0,0.0,1.0,0.0,0.0

子任务二：推荐系统

1.根据子任务一的结果，计算出与用户 id 为 6708 的用户所购买相 同商品种类最多的前 10 位用户 id（只考虑他俩购买过多少种相 同的商品，不考虑相同的商品买了多少次），并根据 Hive 的 dwd 库中相关表或 MySQL 数据库 shtd\_store 中相关表，获取到这 10 位用户已购买过的商品，并剔除用户 6708已购买的商品，通过计 算这 10 位用户已购买的商品（剔除用户 6708 已购买的商品）与用户 6708已购买的商品数据集中商品的余弦相似度累加再求均值，输出均值前 5 商品 id 作为推荐使用，将执行结果截图粘贴至客户端桌面【Release\任务 C 提交结果.docx】中对应的任务序号下。

结果格式如下：

----------------推荐 Top5 结果如下-----------------

相似度 top1(商品 id：1，平均相似度：0.983456)

相似度 top2(商品 id：71，平均相似度：0.782672)

相似度 top3(商品 id：22，平均相似度：0.7635246)

相似度 top4(商品 id：351，平均相似度：0.7335748)

相似度 top5(商品 id：14，平均相似度：0.522356)

**任务 D：数据采集与实时计算（20 分）**

环境说明：

|  |
| --- |
| 服务端登录地址详见各任务服务端说明。补充说明：各节点可通过 Asbru 工具或 SSH 客户端进行 SSH 访问； Flink 任务在 Yarn 上用 per job 模式（即 Job 分离模式，不采用 Session 模式），方便 Yarn 回收资源。 |

子任务一：实时数据采集

1.在主节点使用 Flume 采集实时数据生成器 10050 端口的 socket 数据，将数据存入到 Kafka 的 Topic 中（Topic 名称为 order，分区数为 4），使用 Kafka自带的消费者消费 order（Topic）中的数 据，将前 2 条数据的结果截图粘贴至客户端桌面【Release\任务D 提交结果.docx】中对应的任务序号下；

2.采用多路复用模式，Flume 接收数据注入 kafka的同时，将数据 备份到 HDFS 目录/user/test/flumebackup 下，将查看备份目录下的第一个文件的前 2 条数据的命令与结果截图粘贴至客户端桌面【Release\任务 D 提交结果.docx】中对应的任务序号下。

子任务二：使用 Flink 处理 Kafka 中的数据

编写 Scala 代码，使用 Flink 消费 Kafka中 Topic 为 order 的数 据并进行相应的数据统计计算（订单信息对应表结构 order\_info,订单详细信息对应表结构 order\_detail，同时计算中使用 order\_info 或 order\_detail 表中 create\_time 或 operate\_time 取两者中值较大 者作为 EventTime，若 operate\_time 为空值或无此列，则使用 create\_time 填充，允许数据延迟 5s）。

1. 使用 Flink 消费 Kafka中的数据，统计商城实时订单实收金额， 将 key 设置成 totalprice 存入 Redis中。使用 redis cli以 get key 方式获取 totalprice 值，将结果截图粘贴至客户端桌面【R e lease\任务 D 提交结果.docx】中对应的任务序号下，需两次截图，第一次截图和第二次截图间隔 1 分钟以上，第一次截图放前面，第二次截图放后面；
2. 在任务 1 进行的同时，使用侧边流，监控若发现 order\_status 字段为退回完成, 将 key 设置成 totalrefundordercount 存入 Redis中，value 存放用户退款消费额。使用 redis cli以 get key 方式获取 totalrefundordercount 值，将结果截图粘贴至客户端桌面【Release\任务 D 提交结果.docx】中对应的任务序号下，需两次截图，第一次截图和第二次截图间隔 1 分钟以上，第一次截图放前面，第二次截图放后面；
3. 在任务 1 进行的同时，使用侧边流，监控若发现 order\_status 字段为取消订单,将数据存入 MySQL 数据库 shtd\_result 的 order\_info 表中，然后在 Linux 的 MySQL 命令行中根据 id 降序排序， 查询列 id、consignee、consignee\_tel、final\_total\_amount、 feight\_fee，查询出前 5 条，将 SQL 语句复制粘贴至客户端桌面 【Release\任务 D 提交结果.docx】中对应的任务序号下，将执行 结果截图粘贴至客户端桌面【Release\任务 D 提交结果.docx】中对应的任务序号下。

**任务 E：数据可视化（15 分）**

环境说明：

|  |
| --- |
| 数据接口地址及接口描述详见各任务服务端说明。 |

子任务一：用柱状图展示消费额最高的省份

编写 Vue 工程代码，根据接口，用柱状图展示 2020 年消费额最高的5 个省份，同时将用于图表展示的数据结构在浏览器的 console 中进行打印输出，将图表可视化结果和浏览器 console 打印结果分别截图 并粘贴至客户端桌面【Release\任务 E 提交结果.docx】中对应的任务序号下。

子任务二：用柱状图展示消费额最低的省份

编写 Vue 工程代码，根据接口，用柱状图展示 2020 年消费额最低的 5 个省份，同时将用于图表展示的数据结构在浏览器的 console 中进行打印输出，将图表可视化结果和浏览器 console 打印结果分别截图 并粘贴至客户端桌面【Release\任务 E 提交结果.docx】中对应的任务序号下。

子任务三：用折线图展示每年上架商品数量变化

编写 Vue 工程代码，根据接口，用折线图展示每年上架商品数量的变化情况，同时将用于图表展示的数据结构在浏览器的 console中进行打印输出，将图表可视化结果和浏览器 console 打印结果分别截图并 粘贴至客户端桌面【Release\任务 E 提交结果.docx】中对应的任务序号下。

子任务四：用条形图展示平均消费额最高的省份

编写 Vue 工程代码，根据接口，用条形图展示 2020 年平均消费额（四舍五入保留两位小数）最高的 5 个省份，同时将用于图表展示的数据结构在浏览器的 console 中进行打印输出，将图表可视化结果和浏览 器 console 打印结果分别截图并粘贴至客户端桌面【Release\任务 E提交结果.docx】中对应的任务序号下。

子任务五：用折柱混合图展示省份平均消费额和地区平均消费额

编写 Vue工程代码，根据接口，用折柱混合图展示 2020 年各省份平 均消费额（四舍五入保留两位小数）和地区平均消费额（四舍五入保留两位小数）的对比情况，柱状图展示平均消费额最高的 5 个省份，折线图展示这 5 个省所在的地区的平均消费额变化，同时将用于图表展示的数据结构在浏览器的 console中进行打印输出，将图表可视化 结果和浏览器 console 打印结果分别截图并粘贴至客户端桌面【Release\任务 E 提交结果.docx】中对应的任务序号下。

**任务 F：综合分析（10 分）**

子任务一：F link 有哪些重启策略？各个重启策略如何配置？

在任务 D中使用到了 Flink，Flink 在运行 job 时可能会出现各种问 题，从而会导致其失败或者重启，对于类似于网络波动造成的运行失 败可以采取相对应重启策略来重试，请问 Flink 有几种重启策略（中文）？分别怎么配置这些重启策略？将内容编写至客户端桌面【Release\任务 F 提交结果.docx】中对应的任务序号下。

子任务二：Hadoop 有哪些类型的调度器？简要说明其工作方法。

简要描述 Hadoop 有哪些类型的调度器并简要说明其工作方法，将内 容编写至客户端桌面【Release\任务 F 提交结果.docx】中对应的任务序号下。

子任务三：分析下一年度的建仓目的地。

根据任务E的图表，分析各省份的经济现状，公司决定挑选 3 个省份进行仓储建设，请问应该在哪些省份建设？将内容编写至客户端桌面【Release\任务 F 提交结果.docx】中对应的任务序号下。