全国职业院校技能大赛
赛项规程

赛项名称：大数据应用开发

英文名称：Big Data Application Development

赛项组别：高等职业教育（师生同赛）

赛项编号：GZ033

一、赛项信息

<table>
<tr><td colspan="4">赛项类别</td></tr>
<tr><td colspan="4">☑每年赛　　□隔年赛（□单数年/□双数年）</td></tr>
<tr><td colspan="4">赛项组别</td></tr>
<tr><td colspan="4">□中等职业教育　☑高等职业教育</td></tr>
<tr><td colspan="4">□学生赛(□个人/□团体)　□教师赛（试点）　☑师生同赛（试点）</td></tr>
<tr><td colspan="4">涉及专业大类、专业类、专业及核心课程</td></tr>
<tr><td>专业大类</td><td>专业类</td><td>专业名称</td><td>核心课程
（对应每个专业，明确涉及的专业核心课程）</td></tr>
<tr><td rowspan="19">51 电子与信息大类</td><td rowspan="19">5102 计算机类</td><td rowspan="6">510205 大数据技术</td><td>数据采集技术</td></tr>
<tr><td>数据预处理技术</td></tr>
<tr><td>大数据分析技术应用</td></tr>
<tr><td>数据可视化技术与应用</td></tr>
<tr><td>数据挖掘应用</td></tr>
<tr><td>大数据平台部署与运维</td></tr>
<tr><td rowspan="5">510201 计算机应用技术</td><td>数据库技术及应用</td></tr>
<tr><td>前端设计与开发</td></tr>
<tr><td>信息采集技术</td></tr>
<tr><td>数据分析方法</td></tr>
<tr><td>系统部署与运维</td></tr>
<tr><td rowspan="3">510202 计算机网络技术</td><td>Linux 操作系统管理</td></tr>
<tr><td>程序设计基础</td></tr>
<tr><td>数据库应用技术</td></tr>
<tr><td rowspan="4">510203 软件技术</td><td>程序设计基础</td></tr>
<tr><td>数据库技术</td></tr>
<tr><td>面向对象程序设计</td></tr>
<tr><td>数据结构</td></tr>
<tr><td>510206 云计算</td><td>Linux 操作系统</td></tr>
</table>

		技术应用	程序设计基础
			数据库技术
			Web 应用开发
		510209 人工智能技术应用	程序设计基础
			Linux 操作系统
			数据库技术
			人工智能数据服务
		510211 工业互联网技术	程序设计基础
			数据库应用基础
			工业互联网数据采集技术
			工业互联网数据分析技术
			数据采集与处理

对接产业行业、对应岗位（群）及核心能力

产业行业	岗位（群）	核心能力 （对应每个岗位（群），明确核心能力要求）
战略性新兴产业-新一代信息技术	大数据实施与运维	大数据平台搭建部署与基本使用，以及大数据集群运维
		大数据平台管理、大数据技术服务
	数据分析处理	分析用户业务需求，制订大数据项目解决方案
		开发数据采集、抽取、清洗、转换与加载等数据预处理模型
		基于行业应用与典型工作场景，解决业务需求
	大数据分析与可视化	安装部署与使用数据分析工具，运用大数据分析平台完成大数据分析任务
		数据可视化设计，开发应用程序进行数据可视化展示，撰写数据可视化结果分析报告
		基于行业应用与典型工作场景，解决业务需求
	程序设计	数据库应用、前端开发等程序设计能力
	数据采集与分析	数据采集、使用工具进行数据分析
	信息系统运行维护	信息系统部署与运维

二、竞赛目标

“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。

2、赛项模块、比赛时长及分值配比如下：

表 3-2　赛项模块比赛时长及分值配比

模块		主要内容	比赛时长	分值
模块一	大数据应用开发	竞赛以电商大数据及工业大数据为业务背景，主要设置以下竞赛任务： **任务 A：大数据平台搭建（容器环境）** 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。 **任务 B：离线数据处理** 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。 **任务 C：数据挖掘** 运用常用的机器学习方法对数据进行数据挖掘分析。 **任务 D：数据采集与实时计算** 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。 **任务 E：数据可视化** 基于前端框架 Vue.js 和后端 REST	8 小时	100 分

		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。 **任务 F：综合分析** 对大数据技术的业务分析、技术分析及报告撰写能力。 **任务 G：职业素养** 综合职业素养，包括团队分工明确合理、操作规范、文明竞赛等内容。		

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

（一）竞赛时间表

表 5-1　竞赛时间

日期	时间	内容
竞赛前两日	18:00 之前	裁判报到
	19:00—20:00	裁判工作会议
竞赛前一日	12:00 之前	各参赛队报到
	10:00—11:00	工作人员（含监考）培训会
	15:30—16:00	赛前领队会
	16:00—16:30	参赛队熟悉比赛场地
	17:00—18:00	现场裁判赛前检查，封闭赛场
竞赛当日	07:00—08:00	参赛队集合前往比赛现场
	08:00—08:10	赛场检录
	08:10—08:30	一次加密：参赛队抽取参赛编号
	08:30—08:45	二次加密：参赛队抽取赛位号
	08:45—09:00	参赛队进入比赛赛位，进行赛前软、硬件检查、题目发放
	09:00—17:00	竞赛进行
	17:00—17:20	收取各参赛队赛题及比赛结果文档
	17:00—19:00	申诉受理
	19:00—19:30	三次加密：竞赛结果等文件加密
	19:30—23:00	成绩评定与复核
	23:00—23:30	加密信息解密
	23:30—24:00	成绩汇总及报送
竞赛后一日	08:00—11:00	成绩公布
	11:00—12:00	闭赛式

（二）竞赛流程图

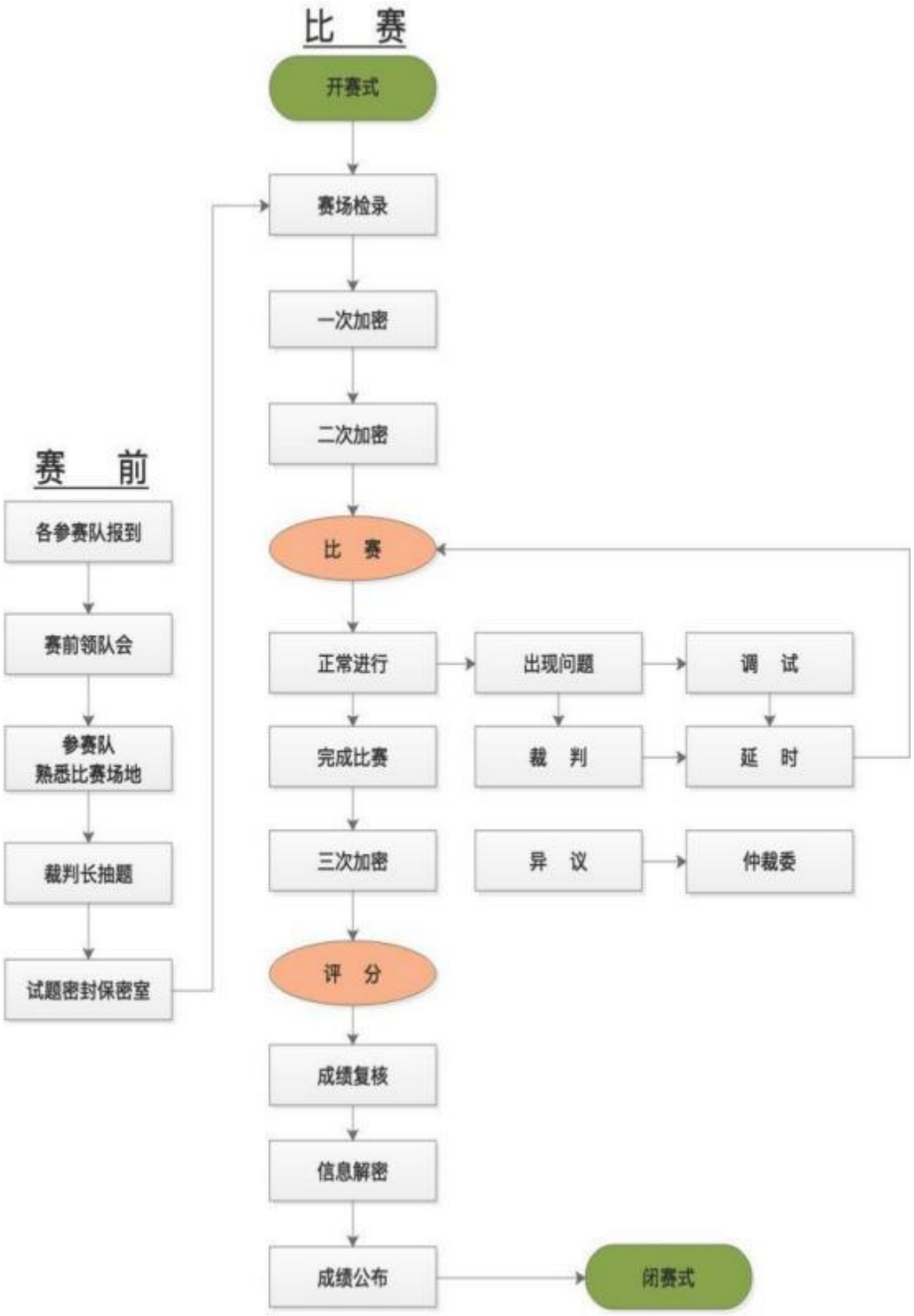

图 5-1　竞赛流程

六、竞赛规则

（一）选手报名：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）熟悉场地：竞赛前 1 日安排各参赛队领队、参赛选手熟悉赛场。

（三）入场规则：参赛选手按规定时间到达指定地点，必须携带参赛证件，进行检录、一次加密、二次加密等流程，最终确定工位，选手迟到 10 分钟取消比赛资格。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。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，如手机、U 盘、移动硬盘等。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。

（四）赛场规则：在比赛过程中，参赛选手如有疑问，应举手示意，现场裁判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，现场裁判、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，经裁判长确认，予以启用备用设备。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。

（五）离场规则： 比赛时间结束，选手应全体起立，结束操作。参赛选手要确认已成功提交竞赛要求的文档，裁判员与参赛选手一起签字确认，经工作人员查收清点所有文档后方可离开赛场，离开赛场

时不得带走任何资料。

（六）成绩评定与结果公布：比赛结束，经加密裁判对各参赛选手提交的竞赛结果进行第三次加密后，评分裁判方可入场进行成绩评判。最终竞赛成绩经复核无误，由裁判长、监督仲裁长签字确认后，以纸质形式向全体参赛队进行公布，并在闭赛式上予以宣布。

（七）其它未尽事宜，将在赛前向各领队做详细说明。

七、技术规范

本赛项引用的国际、国家、行业技术、职业资格标准与规范如下：

表 7-1　基础标准

标准号/规范简称	名称
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 12991.1-2008	信息技术 数据库语言 SQL 第 1 部分：框架
GB/Z 21025-2007	XML 使用指南
GB/T 28821-2012	关系数据管理系统技术要求
LD/T 81.1-2006	职业技能实训和鉴定设备通用技术规范

表 7-2　大数据技术相关标准

标准号/规范简称	名称
GB/T 35295-2017	信息技术 大数据 术语
GB/T 37721-2019	信息技术 大数据分析系统功能要求
GB/T 37722-2019	信息技术 大数据存储与处理系统功能要求
GB/T 38672-2020	信息技术 大数据 接口基本要求
GB/T 38673-2020	信息技术 大数据 大数据系统基本要求
GB/T 38675-2020	信息技术 大数据计算系统通用要求
GB/T 38633-2020	信息技术 大数据 系统运维和管理功能要求
GB/T 41778-2022	信息技术 工业大数据 术语
GB/T 41818-2022	信息技术 大数据 面向分析的数据存储与检索技术要求

表 7-3　软件开发与软件工程相关标准

标准号/规范简称	名称
GB/T 14079-1993	软件维护指南
GB/T 15853-1995	软件支持环境
GB/T 17544-1998	信息技术软件包质量要求和测试
GB/T 8566-2007	信息技术 软件生存周期过程
GB/T 22032-2021	系统与软件工程 系统生存周期过程

八、技术环境

（一）竞赛场地

竞赛现场设置竞赛区、裁判区、技术支持区、服务区等。

1. 竞赛区域：每个竞赛工位设工位编号，面积在 9 ㎡左右，工位之间由隔板隔开，确保互不干扰。

2. 裁判区：供裁判工作及休息，配备满足需要的办公设备。

3. 技术支持区：供技术支持人员工作及休息，为竞赛提供技术支持。

4. 服务区：提供医疗等服务保障。

（二）技术平台

1. 竞赛设备

表 8-1 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	大数据赛训管理系统（四合天地大数据	每组 1 套	该系统基于主流云原生技术、大数据技术构建，旨在为学生提供快捷、便利的大数据集群操作环境，帮助他们更好地掌握大数据相关技术和应用。该系统应基于微服务构建，以经典的微服务分层方式划分不同的服务层级，

	实训管理系统V2.0）		利用图形化的工作负载编辑模式快速进行系统的部署和服务管理，有效展示各服务的容器信息，方便实时进行系统运维。系统能够构建大数据平台搭建、数据处理、数据分析、数据可视化等教学实训模块，快速开展教学、实训及竞赛活动，系统应能够生成命令行、桌面级容器环境，可通过不同模式进行访问，方便学生进行集群调试和代码开发。通过使用该系统，学生可以深入了解大数据技术的核心思想和应用场景，增强自己的数据分析和处理能力，提高对数据的认识和运用水平。系统应支持模拟竞赛全业务流程，提供大数据竞赛操作环境。
3	PC 机	每组 4 台	CPU：i5 及以上 内存：不少于 16GB 硬盘：不少于 500GB 显示器：1920*1080 及以上
4	交换机	每组 1 台	8 口及以上千兆交换机

2.软件环境

表 8-2　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

九、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 ma

ster、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 Master 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.apache.spark.examples.SparkPi $SPARK_HOME/examples/jars/spark-examples_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 master 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、order_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 ods.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 partitions 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 ods 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 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_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、d

wd_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 执行 show partitions dwd.dim_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 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_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 查询表 dim_sku_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_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_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、dw

d_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_region 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_user 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_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_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	int	省份表主键	
provincename	text	省份名称	
regionid	int	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	int	订单总数	当月订单总数
year	int	年	订单产生的年
month	int	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 provinceeverymonth 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
provinceid	int	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	int	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	int	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（order_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【Release\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 3	若属于该分类级别 3，

		2	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【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；

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 ord

er_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：Flink 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。

十、赛项安全

（一）比赛环境

1. 赛场的布置，赛场内的器材、设备，应符合国家有关安全规定。

2. 赛场周围要设立警戒线，防止无关人员进入和发生意外事件。

3. 承办院校应提供保障应急预案实施的条件，明确制度和预案。

4. 赛项执委会须会同承办院校制定开放赛场和体验区的人员疏导方案。

5. 大赛期间，赛项承办院校须在赛场设置医疗医护工作站。

6. 参赛选手、赛项裁判、工作人员严禁携带通讯、摄录设备和未经许可的记录用具进入比赛区域。

（二）生活环境

1. 比赛期间，原则上由赛项承办院校统一安排参赛选手食宿。

承办院校须尊重少数民族参赛人员的宗教信仰及文化习俗。

2. 比赛期间安排的住宿场所应具有旅游业经营许可资质。

3. 大赛期间组织的参观和观摩活动的交通安全由赛区组委会负责。

（三）组队责任

1. 各省、自治区、直辖市在组织参赛队时，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。

十一、成绩评定

（一）评分标准

表 11-1　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，包括全量数据抽取、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	正确使用 Vue.js 框架，结合 ECharts 绘	2

	示每年上架商品数量变化	制折线图	
	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

（二）评分方式

1. 裁判人数和组成条件要求

- 本竞赛参与赛项成绩管理的组织机构包括裁判组、监督仲裁组。裁判组实行“裁判长负责制”，设裁判长 1 名、加密裁判 3 名、现场裁判 8 名、评分裁判 16 名，共计 28 人，要求如下：

表 11-2　裁判要求

职责	专业技术方向	知识能力要求	执裁、教学、工作经历	专业技术职称（职业资格等级）	人数
裁判长	电子与信息	信息技术	执裁过全国职业院校技能大赛，教授过信息技术相关课程	高级职称	1
现场裁判	电子与信息	信息技术	执裁过省级竞赛，教授过信息技术相关课程	高级职称	8
评分裁判	电子与信息	信息技术	执裁过省级竞赛，教授过信息技术相关课程	高级职称	16
加密裁判	无	无	无	高级职称	3

- 监督仲裁组对裁判组的工作进行全程监督，并对竞赛成绩抽检复核。

- 监督仲裁组负责接受由参赛队领队提出的对裁判结果的书面申诉，组织复议并及时反馈复议结果。

- 竞赛将制定裁判遴选管理办法、赛事保密细则和预案、命题 管理办法等制度，保证竞赛的公平公正。

2. 裁判评分方法

本赛项采取三次加密，第一次加密裁判组织参赛选手第一次抽签，抽取参赛编号，替代选手参赛证等个人信息；第二次加密裁判 组织参赛选手进行第二次抽签，确定赛位号，替换选手参赛

编号；第三次加密裁判对各参赛队竞赛结果进行加密，替换赛位号。每个环节结束后，数据立即封存于承办校保密室保险柜内，加密裁判直接隔离，在评分结束后进行解密并统计成绩。本赛项采用结果评分，所有任务均为客观评分。根据评分标准设计评分表，对照参考答案和选手提交结果进行评分，并在评分表中进行统计汇总。裁判需进行随机抽签分组，各裁判小组采取随机抽签针对不同任务独立评分，确保成绩评定严谨、客观、准确。

3. 成绩产生方法

各裁判小组完成本组评分后汇总本组评分表，核对成绩，本组裁判成员签字确认后交予裁判长，裁判长汇总各小组的各任务评分表，核对成绩，最终得出竞赛成绩。

4. 成绩审核方法

为保障成绩评判的准确性，监督仲裁组将对赛项总成绩排名前30%的所有参赛队的成绩进行复核；对其余成绩进行抽检复核，抽检覆盖率不得低于 15%。如发现成绩错误以书面方式及时告知裁判长，由裁判长更正成绩并签字确认。复核、抽检错误率超过 5%的，裁判组将对所有成绩进行复核。

5. 成绩公布方法

最终竞赛成绩经复核无误，加密裁判在监督人员监督下进行三次解密，解密后由裁判长、监督仲裁长签字确认，以纸质形式向全体参赛队进行公布，并在闭赛式上予以宣布。

十二、奖项设置

本赛项奖项设团体奖。设奖比例为：以赛项实际参赛队总数

为基数，一、二、三等奖获奖比例分别为 10%、20%、30%（小数点后四舍五入）。

如出现参赛队总分相同情况，按照任务分值权重顺序的得分高低排序，即总成绩相同的情况下比较任务 C 的成绩，任务 C 成绩高的排名优先，如果任务 C 成绩也相同，则按任务 D、任务 B、任务 A、任务 E、任务 F 的成绩进行排名，以此类推完成相同成绩的排序。如果所有任务分值相同，则查看文档撰写规范、职业素养的分值进行排序。

十三、赛项预案

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、客户机：现场提供占总参赛队伍 10%的备用服务器、客户机。

现场应急预案详情，如下：

（1）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。

（2）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。

（3）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。

十四、竞赛须知

（一）参赛队须知

1. 参赛队应该参加赛项承办单位组织的闭赛式等各项赛事活动。

2. 在赛事期间，领队及参赛队其他成员不得私自接触裁判，凡发现有弄虚作假者，取消其参赛资格，成绩无效。

3. 所有参赛人员须按照赛项规程要求按照完成赛项评价工作。

4. 对于有碍比赛公正和比赛正常进行的参赛队，视其情节轻重，按照《全国职业院校技能大赛奖惩办法》给予警告、取消比赛成绩、通报批评等处理。其中，对于比赛过程及有关活动造成重大影响的，以适当方式通告参赛院校或其所属地区的教育行政主管部门依据有关规定给予行政或纪律处分，同时停止该院校参加全国职业院校技能大赛 1 年（届）。涉及刑事犯罪的移交司法机关处理。

5. 各省、自治区、直辖市在组织参赛队时，须安排为参赛队购买大赛期间的人身意外伤害保险。

（二）领队须知

1. 各参赛代表队要发扬良好道德风尚，听从指挥，服从裁判，不弄虚作假。如发现弄虚作假者，取消参赛资格，名次无效。

2. 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。

3. 竞赛过程中，除参加竞赛的选手、执行裁判员、现场工作人员和经批准的人员外，领队、其他人员一律不得进入竞赛现场。

4. 参赛代表队若对竞赛过程有异议，在规定的时间内由领队向赛项仲裁工作组提出书面报告。

5. 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。

6. 领队应及时查看大赛专用网页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. 参赛选手在报名获得确认后，原则上不再更换。如在筹备过程中，选手因故不能参赛，需出具书面说明并按相关参赛选手资格要求补充人员并接受审核；竞赛开始后，参赛队不得更换参赛选手，允许队员缺席。

2. 参赛选手严格遵守赛场规章、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。

3. 参赛选手凭证进入赛场，在赛场内操作期间应当始终佩戴参赛凭证以备检查。

4. 参赛选手进入赛场，不允许携带任何书籍和其他纸质资料（相关技术资料的电子文档工作人员提供），不允许携带通讯工具和存储设备（如 U 盘）。竞赛统一提供计算机以及应用软件。

5. 各参赛队应在竞赛开始前一天规定的时间段进入赛场熟悉环境。入场后，赛场工作人员与参赛选手共同确认操作条件及设备状况，参赛队员必须确认材料、工具等。

6. 竞赛时，在收到开赛信号前不得启动操作设备。在指定赛位上完成竞赛项目，严禁作弊行为。

7. 竞赛过程中，因严重操作失误或安全事故不能进行比赛的，现场裁判员有权中止该队比赛。

8. 选手在比赛期间不能离场，食品、饮水等由赛场统一提供。选手休息、饮食或入厕时间均计算在比赛时间内。

9. 凡在竞赛期间提前离开的选手，当天不得返回赛场。

10. 为培养技术技能人才的工作风格，在参赛期间，选手应当注意保持工作环境及设备摆放符合企业生产“5S”（即整理、整顿、清扫、清洁和素养）的原则，如果过于脏乱，裁判员有权酌情扣分。

11. 在竞赛中如遇非人为因素造成的设备故障，经裁判员确认后，可向裁判长申请补足排除故障的时间。

12. 参赛选手欲提前结束比赛，应向裁判员举手示意，由裁判员记录竞赛终止时间。竞赛终止后，不得再进行任何与竞赛有关的操作。

13. 各竞赛队按照大赛要求和赛题要求提交竞赛结果，禁止在竞赛结果上做任何与竞赛无关的记号。

14. 竞赛操作结束后，参赛队要确认成功提交竞赛要求的文件，裁判员在比赛结果的规定位置做标记，并与参赛队一起签字确认。

15. 仪容仪表：参赛选手需以端庄的仪容仪表，优雅的行为举止，标准规范的操作进行竞赛。

（四）工作人员须知

1. 赛项全体工作人员必须服从统一指挥，要以高度负责的态度做好比赛服务工作。

2. 全体工作人员要按照工作分区准时到岗，尽职尽责，做好职责工作并做好临时性工作，保证比赛顺利进行。

3. 全体工作人员必须佩戴标志，认真检查证件，经核对无误后方可允许相关人员进入指定地点。

4. 如遇突发事件要及时报告，同时做好疏导工作，避免重大事故发生，确保大赛圆满成功。

5. 各工作组负责人，要坚守岗位，组织落实本组成员高效率完成各自工作任务，做好监督协调工作。

6. 全体工作人员不得在比赛场内接打电话，以保证赛场设施

的正常工作。

十五、申诉与仲裁

（一）各参赛队对不符合大赛和赛项规程规定的仪器、设备、工装、材料、物件、计算机软硬件、竞赛使用工具、用品，竞赛执裁、赛场管理，以及工作人员的不规范行为等，可向赛项监督仲裁组提出申诉。申诉主体为参赛队领队。参赛队领队可在比赛结束后（选手赛场比赛内容全部完成）2 小时之内向监督仲裁组提出书面申诉。

（二）书面申诉应对申诉事件的现象、发生时间、涉及人员、申诉依据等进行充分、实事求是的叙述，并由领队亲笔签名。非书面申诉不予受理。

（三）赛项仲裁工作组在接到申诉报告后的 2 小时内组织复议，并及时将复议结果以书面形式告知申诉方。申诉方对复议结果仍有异议，可由领队向赛区仲裁委员会提出申诉。赛区仲裁委员会的仲裁结果为最终结果。

（四）仲裁结果由申诉人签收，不能代收，如在约定时间和地点申诉人离开，视为自行放弃申诉。

（五）申诉方可随时提出放弃申诉。

（六）申诉方不得以任何理由采取过激行为扰乱赛场秩序。

十六、竞赛观摩

本赛项应须提供公开观摩区。竞赛环境依据竞赛需求和职业特点设计，在竞赛不被干扰的前提下安全开放部分赛场。现场观

摩应遵守如下纪律：

（一）观摩人员需由赛项执委会批准，佩戴观摩证件在工作人员带领下沿指定路线、在指定区域内到现场观赛。

（二）文明观赛，不得大声喧哗，服从赛场工作人员的指挥，杜绝各种违反赛场秩序的不文明行为。

（三）观摩人员不得进入比赛区域，不可接触设备，同参赛选手、裁判交流，不得传递信息，不得采录竞赛现场数据资料，不得影响比赛的正常进行。

（四）观摩者不可携带手机、平板电脑、智能手表等通讯工具进入赛场，对于各种违反赛场秩序的不文明行为，工作人员有权予以提醒、制止。

十七、竞赛直播

本赛项使用大屏幕实时转播现场实况。

（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。

（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. 从竞赛正式开始后，全程进行赛场实时录像直播。

（三）直播内容

1. 赛项执行委员会安排专人对赛项开闭赛式、比赛过程进行全程直播和录像。

2. 制作参赛选手、领队采访实录，裁判专家点评和企业人士 采访视频资料，突出赛项的技能重点与优势特色。为宣传、仲裁、资源转化提供全面的信息资料。

以上内容通过赛项网站进行公开，提交技能大赛官网。

十八、赛项成果

赛项资源转化工作以提升职业院校学生技能水平、引领职业 学校专业建设和教学改革为宗旨，形成满足职业教育教学需求、体现先进教学模式、反映职业教育先进水平的共享性资源成果。

（一）成果形式

资源转化成果包括基本资源和拓展资源。

1. 基本资源包括：风采展示、技能概要、教学资源等。

2. 拓展资源包括：理论试题库、项目案例库、技能素材库等。

（二）主要内容

1. 基本资源

- 风采展示：制作赛项宣传片、获奖代表队（选手）风采展示片。

- 技能概要：制作赛项技能介绍、技能操作要点、评价指标等材料。

- 教学资源：制作赛课融通教材、在线课程资源、学术交流资料、教学改革模式成果等资源。

3. 拓展资源

- 理论试题库：制作相关理论知识的试题库。

- 项目案例库：制作以企业真实项目为载体的案例库。

- 技能素材库：制作相关实操技术技能的素材库。

（三）方法途径

1. 基本资源

- 风采展示：对赛项整体流程进行摄影摄像，对获奖代表队（选手）进行采访，最终形成可供专业媒体宣传播放的风采展示。

- 技能概要：参考《职业教育专业简介》、赛项所属产业或覆盖行业中的标准与规范，制作技能介绍、技能操作要点、评价指标等材料，对接院校人才培养方案及企业岗位人才需求。

- 教学资源：以活页式教材、一体化教材模式，开发以学生为中心、以项目为纽带、以任务为载体、以工作过程为导向的赛课融通教材；以文本、视频、演示文稿等形式，开发以教授学生能力为目标的在线课程资源；组织参赛队开展学术交流，针对课程体系设计、教学模式改革、备赛训练经验等内容进行分享交流；组织职业院校开展教学改革研讨，分享各院校在人才培养模式、师资团队建设、产教融合实践等方面的教学改革模式成果。

2. 拓展资源

● 理论试题库：针对赛项理论知识进行细分并完成试题库设计，最终以选择、判断、填空等形式表现。

● 项目案例库：选取多个企业真实项目进行案例库设计，最终以项目化实训指导文档、实际操作视频等形式表现。

● 技能素材库：针对赛项技术技能进行细分并完成素材库设计，最终以知识点讲义文档、实训指导文档、教学演示文稿、实际操作视频等形式表现。

（四）目标数量和完成时间

资源转化及开发计划如下所示：

表 18-1 资源转化表

资源名称			表现形式	资源数量	资源要求	完成时间
基本资源	风采展示	赛项宣传片	视频	1 个	15 分钟以上	赛后 30 天内完成
		风采展示片	视频	1 个	10 分钟以上	赛后 30 天内完成
	技能概要	技能介绍	文本文档	1 份	约 10 千字	赛后 90 天内完成
		技能要点	文本文档	1 份		赛后 90 天内完成
		评价指标	文本文档	1 份		赛后 90 天内完成
		赛课融通教材	文本文档	1 项	约 50 千字	赛后 90 天内完成
			文本文档	40 个	约 20 千字	赛后 90 天内完成

	教学资源	在线课程资源	演示文稿	40 个	配套使用演示文稿	赛后 90 天内完成
			教学视频	40 个	配套使用教学视频	赛后 90 天内完成
		学术交流资料	视频	1 个	10 分钟以上	赛后 90 天内完成
		教学改革模式 成果	视频	1 个	10 分钟以上	赛后 90 天内完成
拓展资源	理论试题库		文本文档	7 份	约 10 千字	赛后 90 天内完成
	项目案例库		文本文档	7 份	约 10 千字	赛后 90 天内完成
	素材资源库		文本文档	40 个	约 20 千字	赛后 90 天内完成
			演示文稿	40 个	配套使用演示文稿	赛后 90 天内完成
			教学视频	40 个	配套使用教学视频	赛后 90 天内完成