# 2022 年全国职业院校技能大赛（高职组）

# “大数据技术与应用”赛项规程

## 一、赛项名称

赛项编号：GZ-2022041

赛项名称：大数据技术与应用

英文名称：Big Data Technology and Application

赛项组别：高职

赛项归属：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。大数据持续激发商业模式创新，不断催生新业态，已成为互联网等新兴领域促进业务创新增值、提升企业核心价值的重要驱动力。

本赛项旨在落实国家“建设数字中国”战略，深化产教融合，协同推动大数据产业创新与发展，大力推进大数据专业技术技能人才培养，赋能经济社会和现代职业教育高质量发展。结合当前大数据产业中的新技术、新要求，通过大赛让参赛选手熟悉一个真实企业级大数据项目中各个环节的实现过程。通过竞赛来检验教学水平，引领和促进职业教育教学改革，促进与产业主流技术技能接轨，营造崇尚技能的社会氛围。

通过大赛培养参赛选手在企业真实项目环境下进行大数据平台搭建（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

可视化以及综合分析等方面的能力；同时培养选手的理解力、沟通力、抗压力、6S 规范等职业素质；激发学生自主学习能力和解决问题的能力，达到“以赛促教、以赛促学、以赛促改、赛课融通、赛训结合”目的。

赛项围绕大数据产业相关岗位的实际需求和要求进行设计，通过大赛搭建校企合作的平台，深化产教融合，推进产教融合人才培养模式，提升大数据技术与应用专业及其他相关专业毕业生的就业竞争能力，同时强化竞赛成果转化，促进相关教材、资源、师资、认证、实习就业等全面建设，推动院校和企业联合培养大数据人才，加强学校教育与产业发展的有效衔接，促进职业院校电子与信息大类相关专业共同发展，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

（一）选手需具备能力

本赛项基于企业真实项目和工作模块，结合高职大数据技术与应用专业教学标准，充分考量企业岗位对学生职业技能的最新需求，在规定的时间内完成指定大数据模块。其中，主要考核参赛选手在大数据平台搭建（容器环境）、离线数据处理、数据挖掘、数据采集与实时计算、数据可视化以及综合分析等方面的技能。此外，竞赛同时考核参赛选手工作组织和团队协作能力、沟通和人际交往能力、解决问题能力以及致力于紧跟行业发展步伐的自我学习能力。

本项目竞赛内容通过对技能实操表现来评估知识理解以及技能掌握的熟练程度，将不再另外举行知识及理解性质的理论测试。

（二）竞赛模块

1. 竞赛时间

竞赛总时长为 8 小时。各参赛队在规定的时间内，独立完成“竞赛

内容”规定的竞赛模块。

2.竞赛内容

本竞赛结合国内行业、企业的实际业务模型；本竞赛只考核技能部分，不涉及理论。本竞赛进行的技能实操考核，涉及大数据平台搭建（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。

表 3-1 竞赛内容

| 序号 | 比赛模块 | 分数占比 | 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 数据仓库等的综合应用能力，使用 Java、Scala 等开发语言，完成离线数据抽取、数据清洗、数据指标统计等操作 |
| 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 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 |

1.各任务模块的分值比例参考上表，各任务模块包含的子任务分值由专家组命题时确定。

2.关于最终赛题将由专家组讨论决定。其中，各模块的详细内容描述如下：

（1）大数据平台搭建（容器环境）

依据大数据平台的技术特点能够独立解压、安装、配置。对不同的组件进行文件参数配置，日志查看、状态查看、服务启动、组件部署等。

参赛选手需要掌握以下并不仅限于以下技能：

- Docker 容器基础操作；
- Hadoop 伪分布式安装配置；
- Hadoop 完全分布式安装配置；
- Hadoop HA 安装配置；
- Spark 安装配置（Standalone 模式）；
- 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；
- Sqoop 安装配置。

（2）离线数据处理

利用 Scala、Java 等开发语言，对关系型数据库中的离线存量数据进行全量数据抽取、增量数据抽取，将数据存入 Hive 数据仓库，完成数据清洗、数据转化以及相关的数据指标计算等工作。

参赛选手需要掌握以下并不仅限于以下技能：

- Java 项目工程创建与配置；
- Java 应用开发；
- Scala 项目工程创建与配置；
- Scala 应用开发；
- Pom 文件配置；

- Maven 本地仓库配置使用；
- 基于 Sqoop 的数据处理方法；
- 基于 MapReduce 的数据清洗处理方法；
- 基于 Spark 的数据清洗处理方法；
- 基于 Hive 的数据清洗处理方法；
- 数据仓库基本架构及概念；
- 数据仓库星型模型；
- 数据仓库雪花模型。

（3）数据挖掘

利用 Scala 开发语言，基于 Spark ML 机器学习库，根据既有数据完成数据处理建立数据模型完成数据分析、数据挖掘操作。

参赛选手需要掌握以下并不仅限于以下技能：

- Scala 应用开发；
- 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；
- 回归模型应用；
- 聚类模型应用；
- 决策树模型应用；
- 随机森林模型应用。

（4）数据采集与实时计算

基于 Flume、Kafka 组件对实时数据进行采集传输，利用 Scala 开发语言，使用 Flink 对消费实时数据进行相关的数据指标计算等工作。

参赛选手需要掌握以下并不仅限于以下技能：

- Scala 项目工程创建；

- Scala 应用开发
- Pom 文件配置；
- Maven 本地仓库配置使用；
- Redis 基本操作；
- 基于 Flume 及 Kafka 的数据采集方法；
- 基于 Flink 的实时数据处理方法。

（5）数据可视化

对数据进行可视化展示，结合后端 REST 风格的数据接口，利用前端框架 Vue.js 以及数据可视化图表组件 ECharts，将数据分析结果以柱状图、饼图、条形图等图表进行展示。

参赛选手需要掌握以下并不仅限于以下技能：

- Vue.js 项目工程创建；
- Vue.js 框架应用开发；
- ECharts 组件应用开发；
- 根据需求使用 ECharts 绘制柱状图；
- 根据需求使用 ECharts 绘制折线图；
- 根据需求使用 ECharts 绘制折柱混合图；
- 根据需求使用 ECharts 绘制玫瑰图；
- 根据需求使用 ECharts 绘制气泡图；
- 根据需求使用 ECharts 绘制饼状图；
- 根据需求使用 ECharts 绘制条形图；
- 根据需求使用 ECharts 绘制雷达图；
- 根据需求使用 ECharts 绘制散点图；

（6）综合分析

依据数据挖掘分析结果，在综合理解业务数据的基础上，根据题目

要求进行分析，并编写输出分析报告。

参赛选手需要掌握以下并不仅限于以下技能：

- 根据要求结合回归算法结果，说明聚类对业务发展的用途及经营策略影响；
- 根据要求结合聚类算法结果，说明聚类对业务发展的用途及经营策略影响；
- 根据要求结合决策树算法结果，说明决策树对业务发展的用途及经营策略影响；
- 根据要求结合随机森林算法结果，说明随机森林对业务发展的用途及经营策略影响；
- 根据要求结合竞赛过程，对过程中的相关问题提出合理化建议及解决方案。

## 四、竞赛方式

1. 每支参赛队由 3 名选手组成。团体赛参赛队可配指导教师，指导教师须为本校专兼职教师，每队限报 2 名指导教师，竞赛期间不允许指导教师进入赛场进行现场指导。参赛选手和指导教师报名获得确认后不得随意更换；

2. 本赛项设单一场次，所有参赛队在现场根据给定的项目模块，在 8 小时内相互配合，采用小组合作的形式完成赛项模块，最后以提交的截图和文档作为最终评分依据；

3. 不计参赛选手的个人成绩，统计竞赛队的总成绩进行排序。

## 五、竞赛流程

（一）竞赛流程图

2022 年大数据技术与应用赛项的竞赛流程如图 5-1 所示。

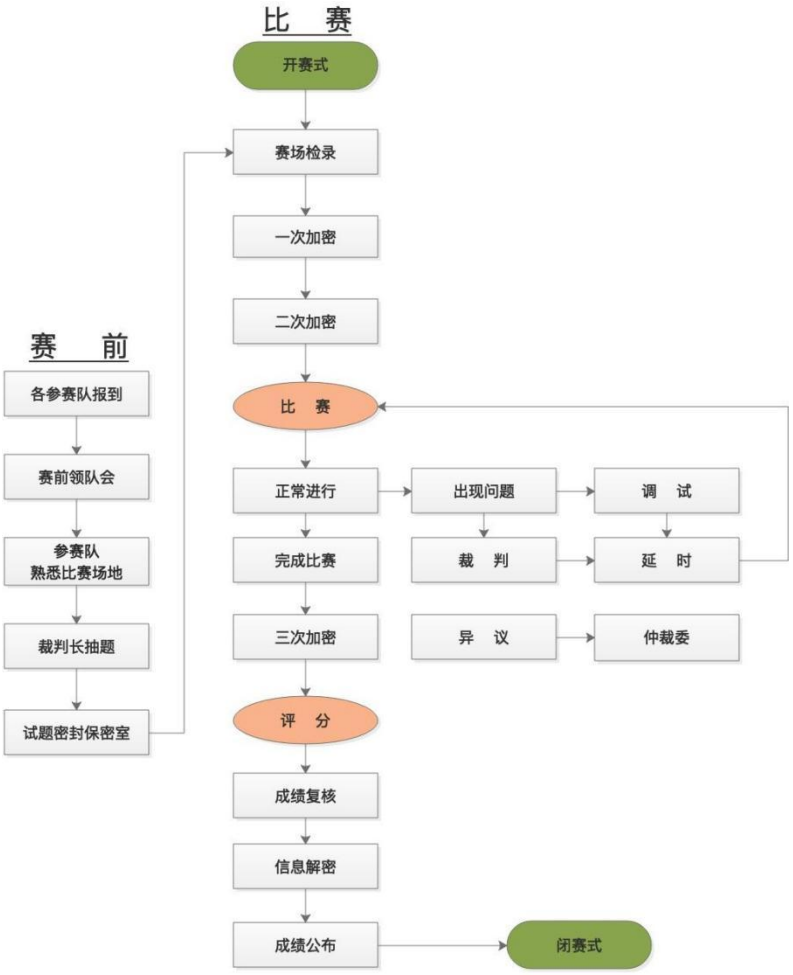


图 5-1 竞赛流程图

## （二）竞赛时间表

表 5-1 竞赛时间表

| 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 |
| | 20:00 之前 | 各参赛队报到 |
| 竞赛前一日 | 10:00—11:00 | 工作人员（含监考）培训会 |
| | 15:00—15:30 | 开幕式 |
| | 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 | 比赛 |
| | 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，报送及公布 |
| 竞赛后一日 | 09:00—10:00 | 闭赛式 |

## 六、竞赛赛卷

（一）专家组建立赛卷库

本赛项建立竞赛赛卷库，样题由全国职业院校技能大赛执委会组织专家组完成，并基于全国职业院校技能大赛相关文件和相关教学标准、职业标准要求，完成竞赛赛卷库建设。制作完成的竞赛赛卷库于开赛前1个月，通过大赛信息发布平台www.chinaskills-jsw.org公开。其中，竞赛样卷与竞赛规程同步发布。

（二）裁判长确定赛题

基于已经公布的竞赛赛卷库，赛前三天内在监督仲裁长监督下裁判长指定相关人员抽取其中2套赛卷（A卷为竞赛用赛卷、B卷为备用赛卷）。专家组将A、B赛卷中不超过30%的内容进行重新编制，并封存于承办院校保密室中。保密室全程监控，并安排专人把守。

比赛完成后，包括参赛选手在内的任何人，都不得将赛题带离赛场，由现场裁判对赛题进行回收。

赛卷样式具体参考样卷，见附件。

## 七、竞赛规则

1. 参赛选手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。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。参赛选手的资格审查工作按照《全国职业院校技能大赛制度汇编》要求

执行。

2. 比赛工位通过抽签决定，比赛期间参赛选手原则上不得离开比赛场地。参赛选手按规定时间到达指定地点，凭参赛相关凭据进入赛场。选手迟到 10 分钟取消比赛资格。

3. 竞赛所需的硬件、软件和辅助工具统一提供，选手不得私自携带任何移动存储、辅助工具、移动通信等设备进入赛场。

4. 参赛选手在赛前 15 分钟进入比赛工位，并由队长领取比赛信息。比赛正式开始后方可进行相关操作。如出现较严重的违规、违纪、舞弊等现象，经裁判组裁定取消比赛成绩。

5. 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。

6. 比赛时间结束，选手应全体起立，结束操作。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。

7. 赛项裁判应严格遵守赛项各项规章制度，确保比赛公平、公正、公开。比赛当天 8:00 起，赛项裁判应上交所有通信设备，由赛项执委会统一保管，并安排赛项裁判在指定区域休息或工作，直至赛项成绩评定结束。

8. 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。

9. 本赛项各参赛队最终成绩，由承办单位信息员在监督仲裁组监督下录入赛务管理系统。承办单位信息员对成绩数据审核后，将赛务系统中录入的成绩导出打印，经赛项裁判长审核无误后，签字。

承办单位信息员将裁判长确认的电子版赛项成绩上传赛务管理系统；同时，将裁判长签字的纸质打印成绩单报送大赛执委会。

10. 赛项结束后，专家工作组根据裁判判分情况，分析参赛选手在比赛过程中对各知识点、技术的掌握程度，并将分析报告报备大赛执委会办公室，执委会办公室根据实际情况适时公布。

11. 赛项中每个比赛环节裁判判分的原始材料和最终成绩等结果性材料，经监督仲裁组人员和裁判长签字后，装袋密封留档；并由赛项承办院校封存，委派专人妥善保管。

12. 其它未尽事宜，将在赛前向各领队做详细说明。

## 八、竞赛环境

（一）赛场布局要求

竞赛现场设置场内竞赛区、裁判工作区、技术支持区、服务区等。

1. 场内竞赛区域。每个竞赛工位标有醒目的工位编号，每个工位面积在 9 ㎡左右，工位之间由隔板隔开，确保参赛队之间互不干扰。赛场要求竞赛过程全程无死角视频监控，监控录像保存 3 个月。环境标准要求保证赛场采光（大于 500lux）、照明和通风良好；提供稳定的水、电，并提供应急的备用电源；提供足够的干粉灭火器材。

2. 裁判工作区。供裁判休息及工作场地。共配有电脑 15 台；A4 激光打印机 1 台；桌椅 15 套；饮水机、纸杯、文具用品若干。

3. 技术支持区。为技术支持人员的工作场地，为参赛选手竞赛提供技术支持。

4. 服务区。提供医疗等服务保障，并用隔离带隔离。

（二）赛事安全要求

1. 禁止选手及所有参加赛事的人员，携带任何有毒有害物品进入竞赛现场。场内竞赛区为参赛队提供统一的竞赛设备，无需选手自带任何工具及附件。

2. 承办单位应设置专门的安全防卫组，负责竞赛期间健康和安全事务。主要包括检查竞赛场地、与会人员居住地、车辆交通及其周围环境的安全防卫；制定紧急应对方案；监督与会人员食品安全与卫生；分析和处理安全突发事件等工作。

3. 赛场须配备相应医疗人员和急救人员，并备有相应急救设施。

4. 承办方应按照疫情防范要求做好赛场各项工作，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，赛场地面张贴荧光疏散指示箭头，赛场出入口专人负责，随时保证安全通道的畅通无阻。

## 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。

表 9-1 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 12991-2008 | 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 28821-1012 | 关系数据管理系统技术要求 |
| LD/T81.1-2006 | 职业技能实训和鉴定设备技术规范 |

表 9-2 大数据技术相关标准

| 标准号/规范简称 | 名称 |
|---|---|

| | |
|---|---|
| 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 38676-2020 | 信息技术 大数据存储与处理系统功能测试要求 |
| GB/T 38643-2020 | 信息技术 大数据分析系统功能测试要求 |
| 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 |

表 9-3 软件开发标准

| 标准号/规范简称 | 名称 |
|---|---|
| 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

技术平台软硬件设备组成如表 10-1 所示。

表 10-1 竞赛硬件设备

| 序号 | 设备名称 | 数量 | 备注 |
|---|---|---|---|
| 1 | 服务器 | 每组1台 | CPU：性能不低于 Intel 至强银牌 4208<br>内存：不少于 128GB<br>硬盘：总容量不少于 1TB<br>网卡：千兆 |
| 2 | 大数据竞赛平台（四合天地大数据实训管理系统） | 每组1套 | 1.四合天地大数据实训管理系统内嵌基于 K8S、Docker 引擎的容器云平台，该软件提供镜像上传存储、Docker 镜像复制、私有镜像仓库管理以及镜像权限控制等功能；支持单个或多个 K8S 集群的导入并进行权限控制；支持在名称空间中以微服务方式将工作负载划分到不同分层，可为每一个名称空间自定义布局；支持图形化的工作负载编辑，快速完成对容器的编排；在工作负载中可将 Deployment 的历史版本、所属的 Pod 列表、Pod 的关联事件、容器信息进行有效组织及展示；可对接 NFS、CephFS 等常用存储类型，并且支持对 CephFS 类型存储卷声明执行扩容和快照；容器文件浏览器支持从容器中进行文件的上传和下载；<br>2.系统基于 Linux 系统部署，支持多角色如（管理员、教师、学生）管理、专业管理、班级管理、用户管理、操作日志、系统设置、镜像环境管理、实训管理、课程管理、实训监控、实验环境、在线实训、个人中心等功能；支持通过 VNC、SSH 等多种模式访问竞赛平台；<br>3.支持模拟大数据平台搭建（容器环境）、离线数据处理、数据挖掘、数据采集与实时计算、数据可视化等贯穿 |

| | | | |
|---|---|---|---|
| | | | 大数据技术的相关知识点，提供大数据竞赛所需的在线操作环境，所涉及开发语言包括 Java、Scala、HTML、JavaScript 等。 |
| 3 | PC 机 | 每组 3 台 | 竞赛选手比赛使用。性能相当于 i5 处理器，不小于 8G 内存，不小于 200G 硬盘，显示器要求 1024*768 以上。 |
| 4 | 交换机 | 每组 1 台 | 5 口及以上千兆交换机 |

（二）软件环境

表 10-2 竞赛软件环境

| 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 |
| | 大数据分析平台组件 | Hadoop 2.7.7 |
| | | Yarn 2.7.7 |
| | | ZooKeeper 3.4.6 |
| | | Hive 2.3.4 |
| | | JDK 1.8 |
| | | Flume 1.7.0 |
| | | Sqoop 1.4.2 |
| | | Kafka 2.0.0 |
| | | Spark 2.1.1 |
| | | Flink 1.10.2 |
| | | Redis 4.0.1 |
| | 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 2.11 |
| | | Java 8 |
| | 开发工具 | IDEA 2021 (Community Edition) |
| | | Visual Studio Code 1.69 |
| | SSH 工具 | Asbru-cm 或 Ubuntu SSH 客户端 |
| | 数据库连接工具 | 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）评分原则

本赛项采用结果评分，根据评分标准设计评分表。

1. 评分表样例

评分表按照选手对应题目要求实现过程及结果进行评分，具体评分样表如表 11-1 所示。

表 11-1 评分表样例

| 模块 | 任务 | 主要知识与技能点 | 分值 |
|---|---|---|---|
| 模块 A:大数据平台搭建（容器环境） | 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 |
| | 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 |
| | 任务三：Flink on Yarn 安装配置 | Flink 的解压安装、环境变量配置、运行测试等 | 4 |
| | 小计 | | 15 |
| 模块 B:离线数据处理 | 任务一：数据抽取 | 从 MySQL 中进行离线数据抽取，包括全量数据抽取和增量数据抽取操作 | 10 |
| | 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 | 8 |
| | 任务三：指标计算 | 在 dwd、dws 层对数据进行相关数据指标的统计、计算等操作 | 7 |
| | 小计 | | 25 |
| 模块 C:数据挖掘 | 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 |
| | 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 |
| 模块 D:数据采集与实时计算 | 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 |
| | 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，包括实时数据统计计算、Redis 基本操作、Kafka 基本操作等 | 12 |
| | 小计 | | 20 |
| 模块 E：数据可视化 | 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 | 2 |
| | 任务二：用饼状 | 正确使用 Vue.js 框架，结合 ECharts 绘制 | 3 |

| | | | |
|---|---|---|---|
| | 图展示各地区消费能力 | 饼状图 | |
| | 任务三：用折线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 | 3 |
| | 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 |
| | 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 |
| 模块 F:综合分析 | 任务一：如何解决 Job 运行效率低的问题 | 正确分析如何解决 Job 运行效率低的问题 | 4 |
| | 任务二：对于数据挖掘模块中的用户推荐有什么好的建议 | 正确分析数据挖掘模块中的用户推荐并提供合理化建议 | 3 |
| | 任务三：简要描述任务过程中的问题并进行总结 | 合理描述任务过程中的问题并进行总结 | 3 |
| | 小计 | | 10 |
| 模块 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 |

**注：以上仅为示例，专家组命题时保持各模块总分值不变，各模块所包含子任务的分值专家组命题时可以做适当微调，具体以专家组实际命制的赛题为准。**

2. 三次加密原则

比赛过程采取三次加密，通过抽取参赛编号、工位号和竞赛成果号，屏蔽参赛队信息，每个环节设置一名独立裁判，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，确保成绩评定公平、公正。

3. 独立评分原则

根据裁判分工，负责相同模块评分工作的不同裁判，采取随机抽签独立评分，确保成绩评定严谨、客观、准确。裁判进行随机抽签分组，杜绝主观意愿组队，各自完全独立评分，裁判员间互不干涉，比赛监督人员可随机监督。

4. 错误不传递原则

各模块分别计算得分，错误不传递，按规定比例计入选手总分。

5. 抽查复核原则

（1）为保障成绩评判的准确性，监督仲裁组对赛项总成绩排名前30%的所有参赛队伍（选手）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%。

（2）监督仲裁组需将复检中发现的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。

（3）复核、抽检错误率超过5%的，则认定为非小概率事件，裁判组需对所有成绩进行复核。

（二）评分方法

1. 竞赛满分为100分。最终成绩按100分制进行排名。

2. 竞赛采取三次加密。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。

3. 裁判长正式提交评分结果并复核无误后，加密裁判在监督人员监督下进行三层解密：竞赛结果编号到工位号解密；工位号到参赛编号解密；参赛编号到参赛队名称解密。

4. 为保障成绩评判的准确性，监督仲裁组对赛项总成绩排名前30%的所有参赛队伍的成绩进行复核；其余成绩进行抽检复核，抽检覆盖率不低于15%。

5. 监督仲裁组在复检中发现错误，需以书面形式及时告知裁判长，由裁判长更正成绩并签字确认。如复核、抽检错误率超过5%，裁判组需对所有成绩进行复核。

6. 在竞赛过程中，参赛选手如有不服从裁判裁决、扰乱赛场秩序、舞弊等行为的，由裁判长按照规定扣减相应分数，情节严重的将取消比赛资格，比赛成绩计0分。

（三）裁判要求

表 11-2 裁判要求

| 序号 | 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 |
|---|---|---|---|---|---|
| 1 | 信息技术 | 信息技术大类 | 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 |
| 2 | 信息技术 | 信息技术大类 | 执裁过省级竞赛，教授过信息技术相关课程 | 高级职称 | 9 |
| 3 | 信息技术 | 大数据 | 执裁过省级竞赛，教授过大数据相关课程 | 高级职称 | 25 |
| 4 | 无 | 无 | 无 | 高级职称 | 3 |
| 裁判总人数 | 竞赛设置裁判38人，包括裁判长1人，加密裁判3人，现场裁判9人，评分裁判25人 | | | | |

注意：承办校可根据本校场地实际情况增加现场裁判数量。

## 十二、奖项设定

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，按照模块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模块C的成绩，模块C成绩高的排名优先，如果模块C成绩也相同，则按模块D、模块B、模块A、模块E、模块F的成绩进行排名，以此类推完成相同成绩的排序。如果所有模块分值相同，则查看文档撰写规范、职业素养的分值进行排序。

获得一等奖的参赛队的指导教师授予“优秀指导教师奖”荣誉称号。

## 十三、赛场预案

（一）应急安全预案

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。赛项出现重大安全问题可以停赛，是否停赛由赛区执委会决定。事后，赛区执委会应向大赛执委会报告详细情况。

相关应急预案如表 13-1 所示。

表 13-1 应急预案表

| 突发事件 | 预防措施 | 事件发生后应对措施 |
| --- | --- | --- |
| 参赛选手发病或受伤 | 在各工位张贴安全操作说明。 | 医务人员应采取紧急救护措施，及时进行救治，如病情或伤势严重，应及时送往最近医院进行救治。 |
| 人员发生食物中毒 | 比赛期间指定的住宿/餐饮场地符合国家相关资质要求。并协调地方卫生部门做好检查工作。 | 立即组织对中毒人员进行救治，必要时送往最近医院进行检查治疗。同时对可疑的食品、饮水及其有关原料、工具设备和场所以及可能受污染的区域采取保留、控制措施，组织开展现场调查，迅速查明原因，并及时向大赛执委会报告。 |
| 设备损坏 | 提前一天服务器全部运行；现场划分备份组。 | 参赛选手举手示意后，监考人员计时，裁判确认后更换备机，并由主裁判确定应计入延时时间。 |
| 设备掉电 | 竞赛前技术人员及监考人员检查所有电源插头，确保牢固；电源线尽量绑扎在参赛选手碰不到的地方，如桌子后面等。竞赛前提醒参赛选手注意尽量不要碰到电源，配置文件要随时保存。 | 参赛选手举手示意后，监考人员计时，裁判确认后重启机器，并由主裁判确定应计入延时的时间。 |

| | | |
|---|---|---|
| 现场网络线缆故障 | 现场走线要规范，尽量走暗槽或现场人员接触不到的地方；对主要线路要在走线槽内留有备线。 | 启用备线。 |
| 临时停电 | 赛场需要双路供电和备用发电机，确保单电源故障不会影响比赛 | 供电线路互为备份，如出现故障，切换线路，经裁判长与赛项执委会商议统一延长比赛时间；若双路电源均出现故障，快速启用备用发电机发电，保证比赛正常运行，经裁判长与赛项执委会商议统一延长比赛相应时间。 |

（二）处罚措施

1. 因参赛队伍原因造成重大安全事故的，取消其获奖资格。

2. 参赛队伍有发生重大安全事故隐患，经赛场工作人员提示、警告无效的，可取消其继续比赛的资格。

3. 赛事工作人员违规的，按照相应的制度追究责任。情节恶劣并造成重大安全事故的，由司法机关追究相应法律责任。

## 十四、赛项安全

赛项安全是全国职业院校技能大赛一切工作顺利开展的先决条件，是本赛项筹备和运行工作必须考虑的核心问题。

（一）组织机构

1. 成立由赛项执委会主任为组长的赛项安全保障小组，成员包括承办院校主抓安全的校领导、学生工作处、后勤处、保卫处、合作企业技术工程师等相关人员；

2. 与地方行政、交通、司法、安全、消防、卫生、食品、质检等相关部门建立协调机制，制定应急预案，及时处置突发事件，保证比赛安全进行。

（二）比赛环境

1. 执委会须在赛前组织专人对比赛现场、住宿场所和交通保障进

行考察，并对安全工作提出明确要求。赛场的布置，赛场内的器材、设备，应符合国家有关安全规定。如有必要，也可进行赛场仿真模拟测试，以发现可能出现的问题。承办单位赛前须按照执委会要求排除安全隐患；

2. 严格控制与参赛无关的易燃易爆以及各类危险品进入比赛场地，不许随便携带书包进入赛场；

3. 配备先进的仪器，防止有人利用电磁波干扰比赛秩序。大赛现场需对赛场进行网络安全控制，以免场内外信息交互，充分体现大赛的严肃、公平和公正性；

4. 大赛期间，承办单位须在赛场管理的关键岗位，增加力量，建立安全管理日志，在赛场封闭后至竞赛结束前对所有比赛场地进行监控，并将监控视频保留 3 个月，防止人为损坏大赛设备影响比赛正常进行。

（三）生活条件

1. 比赛期间，原则上由执委会统一安排参赛选手和指导教师食宿。承办单位须尊重少数民族的信仰及文化，根据国家相关的民族政策，安排好少数民族选手和教师的饮食起居；

2. 比赛期间安排的住宿地应具有宾馆/住宿经营许可资质。以学校宿舍作为住宿地的，大赛期间的住宿、卫生、饮食安全等由执委会和提供宿舍的学校共同负责；

3. 各赛项的安全管理，除了可以采取必要的安全隔离措施外，应严格遵守国家相关法律法规，保护个人隐私和人身自由；

4. 赛项所有裁判与参赛队住宿须在不同酒店。在竞赛日当天早 8 点，由竞赛执委会工作人员收缴裁判所有通信设备，直至竞赛成绩发布后再归还裁判；

5. 竞赛期间，除现场裁判外，其余裁判由竞赛执委会统一安排休息场所。在此期间，裁判人员不得随意出入，避免与参赛队代表取得联

系。

（四）组队责任

1. 各学校组织代表队时，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、指导教师进行安全教育；

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。

（五）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。赛项出现重大安全问题可以停赛，是否停赛由赛区执委会决定。事后，赛区执委会应向大赛执委会报告详细情况。

## 十五、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 3 名参赛选手组成，须为同校在籍学生，其中队长 1 名。每支参赛队可配 2 名指导教师，指导教师须为本校专兼职教师。不接受跨校组队，同一学校报名参赛队不超过 1 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. 参赛选手及指导教师在报名获得确认后，原则上不再更换。如在筹备过程中，参赛选手和指导教师因故不能参赛，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，经大赛执委会办

公室核实后予以更换。允许队员缺席比赛；允许指导教师缺席比赛；

5. 参赛队按照大赛赛程安排，凭赛项执委会颁发的参赛证、有效身份证件和学生证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、领队和指导教师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，须为参赛选手购买大赛期间的人身意外伤害保险；

10. 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。涉及刑事犯罪的移交司法机关处理。

（二）指导教师须知

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. 领队和指导教师务必带好有效身份证件，在活动过程中佩戴“指导教师证”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、指导教师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，由领队和指导教师负责以书面形式向大赛监督仲裁组反映，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. 领队和指导老师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、抽签确定竞赛赛位号，不得迟到早退；

3. 参赛选手凭竞赛赛位号进入赛场，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，认真核对竞赛赛位号，在指定位置就座；

5. 参赛选手入场后，迅速确认竞赛环境状况，填写相关确认文件，并由参赛队长确认签字（竞赛赛位号）；

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，经项目裁判长确认，予以启用备用计算机；

7. 赛项任务书及相关资料，均保存在竞赛环境的相关文件夹中。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，将相应文档按要求进行提交；

8. 参赛选手需及时保存竞赛内容。对于因各种原因造成的数据丢失，由参赛选手自行负责；

9. 参赛队所提交的结果不得出现地名、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），现场裁判员有权中止该队比赛；

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；

13. 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；

14. 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；

15. 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；

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. 赛项全体工作人员必须服从执委会统一指挥，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. 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向执委会报告，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。

## 十六、申诉与仲裁

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，超过时效不予受理。申诉

时，应按照规定的程序由参赛队领队向赛项监督仲裁工作组递交书面申诉报告。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. 赛项监督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，可由省（市）领队向赛区监督仲裁委员会提出申诉。赛区监督仲裁委员会的仲裁结果为最终结果；

4. 申诉人不得采取过激行为刁难、攻击工作人员，否则视为放弃申诉；

5. 申诉方可随时提出放弃申诉。

## 十七、竞赛观摩

本赛项应须提供公开观摩区，使用大屏幕实时转播现场实况。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：

1. 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；

2. 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；

3. 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；

4. 观摩者不可携带手机、IPAD 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。

## 十八、竞赛直播

本赛项竞赛时采用全过程录像，在不影响比赛的前提下，全过程、全方位安排现场直播，并设直播观摩区，让所有参赛教师和社会人员等观看比赛。赛后邀请媒体采访优秀选手、优秀指导教师、裁判专家或企业人士，突出赛项的技能重点与优势特色，为大赛宣传、资源转化提供全面的信息资料。视频资料也作为竞赛成果提交赛项执委会，作为竞赛历史材料供后续赛项提高进行参考，竞赛过程可作为教学资料进行资源转换，促进相关专业教学发展。

## 十九、资源转化

2022 年全国职业院校技能大赛大数据技术与应用赛项资源转化工作主要聚焦完善、升级已经开发完成的专业核心课程教学资源包，进一步开展师资培养，创新培训课程内容，建设大数据技术及其相关专业的生产实际教学案例库等工作，同时对产教融合校企合作案例进行总结。

承办校是资源转化的第一责任单位，全面负责资源转化工作。

（一）资源内容

资源转化成果包括基本资源和拓展资源，充分体现本赛项技能考核特点：

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、评价指标等材料按竞赛任务模块制作相关文本文档、操作演示视频。

教学资源：开发和制作“大数据技术与应用”教学资源，开发专业教材、教学课件 PPT、技能实训指导书、实训操作视频等数字化专业教材资源。

2. 拓展资源

制作反映本赛项技能特色，并且适用于各教学与训练环节的多样性

辅助资源。包括：专家点评视频、优秀选手访谈视频、试题库、项目案例库、素材库等拓展性资源。

制作完成的赛项资源经审核后上传至大赛指定的网络信息管理平台：www.chinaskills-jsw.org。

（二）预期成果

1. 风采展示：赛项宣传片、选手采访、指导老师和专家采访等宣传视频。

2. 技能概要：技能介绍、技能要点、评价指标等相关文本文档、操作演示视频。

3. 教学资源：系列相关教材和资源的开发。

4. 扩展资源：包括赛项专家和指导老师点评视频、优秀选手访谈视频、案例库、素材资源库、试题库等拓展性资源。

（三）完成时间

资源转化及开发计划如表 19-1 所示：

表 19-1 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 |
| | 教学资源 | 技能训练指导书 | 文本文档 | 1 份 | 约 10 千字 | 赛后 90 天内完成 |
| | | 技能操作规程 | 文本文档 | 1 份 | 约 10 千字 | 赛后 90 天内完成 |
| 拓展 | 案例库 | | 文本文档 | 7 份 | 约 20 千字 | 赛后 90 天内完成 |

| | | | | | |
|---|---|---|---|---|---|
| 资源 | 素材资源库 | 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（微课） | 40 个 | 配套使用微视频 | 赛后 90 天内完成 |
| | | FLASH 动画 | 约 30 分钟 | 配套使用 FLASH 动画 | 赛后 90 天内完成 |
| | 赛题库 | 文本文档 | 1 套 | 约 30 千字 | 赛后 90 天内完成 |
| | 专家和指导教师点评视频 | 视频 | 1 个 | 高清视频 | 赛后 30 天内完成 |
| | 优秀选手访谈 | 视频 | 1 个 | 高清视频 | 赛后 30 天内完成 |

## 二十、其他

无。

## 附件：样卷

# 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。

# 模块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模块服务端说明。**

**补充说明：**宿主机可通过Asbru工具或SSH客户端进行SSH访问；

相关软件安装包在宿主机的/opt目录下，请选择对应的安装包进行安装，用不到的可忽略；

所有模块中应用命令必须采用绝对路径；

进入Master节点的方式为

docker exec -it master /bin/bash

进入Slave1节点的方式为

docker exec -it slave1 /bin/bash

进入Slave2节点的方式为

docker exec -it slave2 /bin/bash

## 任务一：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、 从宿主机/opt目录下将文件hadoop-2.7.7.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对应报告中；

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“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，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对应报告中；

4、 在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）粘贴至对应报告中；

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对应报告中。

## 任务二：Spark on Yarn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、 从宿主机/opt目录下将文件spark-2.1.1-bin-hadoop2.7.tgz复制到容器master中的/opt/software（若路径不存在，则需新建）中，将Spark包解压到路径/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对应报告中；

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对应报告中；

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_2.11-2.1.1.jar 运行的主类为 org.apache.spark.examples.SparkPi，将运行结果截图粘贴至对应报告中（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_2.11-2.1.1.jar）

## 任务三：Flink on Yarn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、 从宿主机/opt目录下将文件flink-1.10.2-bin-scala_2.12.tgz复制到容器master中的/opt/software（若路径不存在，则需新建）中,将Flink包解压到路径/opt/module中(若路径不存在，则需新建)，将完整解压命令复制粘贴至对应报告中；

2、 修改容器中/etc/profile文件，设置Flink环境变量并使环境变量生效。在容器中/opt目录下运行命令flink --version，将命令与结果截图粘贴至对应报告中；

3、 开启Hadoop集群，在yarn上以per job模式（即Job分离模式，不采用Session模式）运

行 $FLINK_HOME/examples/batch/WordCount.jar，将运行结果最后10行截图粘贴至对应报告中。

示例 ：

```
flink run -m yarn-cluster -p 2 -yjm 2G -ytm 2G
$FLINK_HOME/examples/batch/WordCount.jar
```

# 模块B：离线数据处理（25分）

环境说明：

**服务端登录地址详见各模块服务端说明。**

**补充说明：**各主机可通过Asbru工具或SSH客户端进行SSH访问；

Master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。

**注：该Spark版本无法进行本地调试，请打包上传集群调试。**

## 任务一：数据抽取

使用Scala编写spark工程代码，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，将结果截图粘贴至对应报告中；

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，将结果截图粘贴至对应报告中；

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，将结果截图复制粘贴至对应报告中；

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，将结果截图粘贴至对应报告中；

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，将结果截图粘贴至对应报告中；

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，将结果截图粘贴至对应报告中。

## 任务二：数据清洗

使用Scala编写spark工程代码，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（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，将结果截图粘贴至对应报告中；

2、 抽取ods库sku_info表中昨天的分区（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，将结果截图粘贴至对应报告中；

3、 抽取ods库base_province表中昨天的分区（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，将结果截图粘贴至对应报告中；

4、 抽取ods库base_region表中昨天的分区（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_region最新分区中，查询该分区中数据的条数，将结果截图粘贴至对应报告中;

5、 将ods库中order_info表昨天的分区（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，将结果截图粘贴至对应报告中;

6、 将ods库中order_detail表昨天的分区（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，将结果截图粘贴至对应报告中。

## 任务三：指标计算

使用Scala编写spark工程代码，并计算相关指标。

**注：在指标计算中，不考虑订单信息表中order_status字段的值，将所有订单视为有效订单。计算订单金额或订单总金额时只使用final_total_amount字段。需注意dwd所有的维表取最新的分区。**

1、 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语句与执行结果截图粘贴至对应报告中；

| 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、 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表中（表结构如下），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语句与执行结果截图粘贴至对应报告中；

| 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 |

3、 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、客户主键均为降序排序，查询出前5条，将SQL语句与执行结果截图粘贴至对应报告中。

| 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<br>yyyyMMdd_yyyyMMdd<br>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 |

# 模块C：数据挖掘（10分）

## 环境说明：

**服务端登录地址详见各模块服务端说明。**

**补充说明：**各主机可通过Asbru工具或SSH客户端进行SSH访问；

Master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。

该模块均使用Scala编写，利用Spark相关库完成。

## 任务一：特征工程

剔除订单信息表与订单详细信息表中用户id与商品id不存在现有的维表中的记录，同时建议多利用缓存并充分考虑并行度来优化代码，达到更快的计算效果。

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个相同的商品，不考虑相同的商品买了多少次），将10位用户id进行输出，输出格式如下，将结果截图粘贴至报告中：

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），将结果截图粘贴至报告中。

| 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3 1 | 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 2 | 若属于该分类级别3，则内容为1否则为0 |
| …… | | | |

答案格式如下：

----------第一条数据前10列结果展示为：----------

1.0,0.89,0.72,0.0,0.0,0.0,0.0,1.0,0.0,0.0

## 任务二：推荐系统

1、 根据任务一的结果，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个相同的商品，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位用户已购买商品与该数据集中商品的余弦相似度累加再求均值，输出相似度前5商品id作为推荐使用，将执行结果截图粘贴至对应报告中。

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：0.98)

相似度top2(商品id：71，平均相似度：0.78)

相似度top3(商品id：22，平均相似度：0.76)

相似度top4(商品id：351，平均相似度：0.73)

相似度top5(商品id：14，平均相似度：0.52)

# 模块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模块服务端说明。**

**补充说明：**各主机可通过Asbru工具或SSH客户端进行SSH访问；

请先检查ZooKeeper、Kafka、Redis端口是否已启动，若未启动则各启动命令如下：

ZK启动（netstat -ntlp查看2181端口是否打开）

/usr/zk/zookeeper-3.4.6/bin/zkServer.sh start

Redis启动（netstat -ntlp查看6379端口是否打开）

/usr/redis/bin/redis-server /usr/redis/bin/redis.conf

Kafka启动（netstat -ntlp查看9092端口是否打开）

/opt/kafka/kafka_2.11-2.0.0/bin/kafka-server-start.sh -daemon（空格连接下一行）/opt/kafka/kafka_2.11-2.0.0/config/server.properties

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。

## 任务一：实时数据采集

1、 在Master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对应报告中；

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对应报告中。

## 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（来源类型和

来源编号这两个字段不考虑，所以在实时数据中不会出现），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为空值或无此属性，则使用create_time填充，允许数据延迟5S，订单状态分别为1001:创建订单、1002:支付订单、1003:取消订单、1004:完成订单、1005:申请退回、1006:退回完成。另外对于数据结果展示时，不要采用例如：1.9786518E7的科学计数法）。

1、 使用Flink消费Kafka中的数据，统计商城实时订单实收金额（需要考虑订单状态，若有取消订单、申请退回、退回完成则不计入订单实收金额，其他状态的则累加），将key设置成totalprice存入Redis中。使用redis cli以get key方式获取totalprice值，将结果截图粘贴至对应报告中，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，将结果截图粘贴至对应报告中，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语句与执行结果截图粘贴至对应报告中。

# 模块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模块服务端说明。 |

## 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对应报告中。

## 任务二：用饼状图展示各地区消费能力

编写Vue工程代码，根据接口，用饼状图展示2020年各地区的消费总额占比，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对应报告中。

## 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对应报告中。

## 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对应报告中。

## 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和地区平均消费额的对比情况，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对应报告中。

# 模块F：综合分析（10分）

## 任务一：如何解决Job运行效率低的问题

在模块B中出现某些Job运行时间较长，你认为可能是哪些情况造成？有什么相应的处理办法吗？将内容编写至对应报告中。

## 任务二：对于数据挖掘模块中的用户推荐有什么好的建议

在模块C中使用基于用户的推荐系统思路对用户的相似性进行计算，从而为每个用户推荐商品，你认为可以从哪些方面再进行优化？这种推荐策略对业务的发展会起到什么样的作用？将内容编写至对应报告中。

## 任务三：简要描述任务过程中的问题并进行总结

将内容编写至对应报告中。