河北省高等职业院校
大数据技术与应用技能大赛

零售大数据分析（样题）

任
务
书

参赛队编号_____

第一部分 竞赛须知

一、 竞赛注意事项

1、 参赛选手应严格遵守赛场规章、操作流程和工艺准则，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、 竞赛所需的硬件、软件和辅助工具由组委会统一布置，选手不得私自携带任何电子设备或其他资料、用品等进入赛场；

3、 比赛完成后，软件和赛题请保留在座位上，禁止将比赛所用的所有物品（包括试卷和草纸）带离赛场；

4、 裁判以各参赛队提交的竞赛结果文档为主要评分依据。所有提交的文档必须按照赛题所规定的命名规则命名，不得以任何形式体现参赛院校、姓名、参赛证编号、赛位号等信息，否则取消竞赛成绩；

5、 本次比赛采用统一网络环境比赛，请不要随意更改客户端和竞赛环境的网络地址信息，对于更改客户端信息造成的问题，由参赛选手自行承担比赛损失；

6、 请不要恶意破坏竞赛环境（如修改竞赛环境密码、删除文件），对于恶意破坏竞赛环境的参赛者，组委会根据其行为予以处罚直至取消比赛资格；

7、 比赛中出现各种问题及时向现场裁判举手示意，不要影响其他参赛队比赛；

二、 竞赛选手须知

1、 任务书如出现缺页、字迹不清等问题，请及时向现场裁判示意，并由现场裁判进行更换；

2、 赛项竞赛时长 4 小时；

3、 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；

4、 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；

5、 参赛选手需及时保存工作记录。对于参赛选手自身原因造成的数据丢失，由参赛选手自行负责；

6、 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；

7、 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；

8、 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；

9、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；

第二部分 竞赛环境及注意事项

一、竞赛环境

每组竞赛选手使用三台计算机和一套大数据竞赛环境，竞赛选手依照本竞赛项目的任务内容，完成任务书要求的相关操作与开发任务。

二、竞赛结果文件提交

1、 所有竞赛结果提交文件夹存放在计算机桌面“竞赛文档”文件夹下，竞赛任务结果截图和文件存放在答案模板下。

2、 请务必按照任务书说明文档题目要求内容截取答案/结果（可分段截取），并按顺序粘贴至答案模板中；在计算机桌面创建“竞赛文档”文件夹，并在该目录中创建 word 文件，用于存放答案截图，文件格式为：“XXX-02.docx（XXX 代表赛位号、02 代表任务二）”。答案文档需学生自行创建并按照习题顺序自行排版。

3、 竞赛结果需提交 Word 文件。

4、 将任务成果 Word 文件压缩为一个 XXX.zip（XXX 代表赛位号）文件，并上传至竞赛平台。

三、 注意事项

1、 检查计算机设备、大数据竞赛环境是否能正常使用。检查竞赛所需的各项设备、软件和竞赛材料等；

2、 竞赛过程中请严格按照竞赛任务中的描述，对大数据竞赛环境进行安装配置、操作使用，对于竞赛前大数据竞赛环境内的配置，与竞赛任务有关，请勿修改、删除；

3、 竞赛任务完成后，不要关闭任何设备，不要对计算机设备或大数据竞赛环境进行加密；

第三部分 竞赛任务

背景描述

当今社会，中国零售业所面临的最大挑战就是顾客和市场需求复杂多变，比起人的经验主义来做决策，只有实时的数据分析和反馈才能适应更快的变化。零售的本质离不开人、货、场这三个核心，围绕这三个核心提升运营的效率，也就是线上线下的成功融合。

为了对零售业中经营模式、管理风格、重视程度、资金投入等做出正确的决策，对其进行数据分析必不可少。现选用在业界广泛使用的“Hadoop”工具，来对该零售行业数据进行分析处理。并综合利用 MySQL、MapReduce、Hive、Sqoop、Spark、Echarts 等技术和 Java、Python 语言对数据进行提取、清洗、整理、计算、表达、分析和可视化处理。

作为分析该零售行业的主要技术人员，你们是这次技术方案展示的核心成员，请按照下面步骤完成本次技术展示任务，并提交技术报告，祝你们成功。

任务一：Hadoop 相关组件安装部署（15 分）

一、Hadoop HA 部署

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，具体部署要求如下：

1、 解压安装 JDK 到路径/usr/local/src，并配置环境变量；截取环境变量配置文件截图保存。

2、 创建 ssh 密钥对，实现主节点与从节点的无密码登录；截取主节点登录其中一个从节点的结果。

3、 将 Zookeeper 组件安装到 /usr/local/zookeeper 路径，zookeeper 的数据目录和日志目录分别为/usr/local/zookeeper/data 和 /usr/local/zookeeper/log。

4、 启动节点 action-1 和 action-2 的 Hadoop 的 NameNode 和 ResourceManager。

二、Hive 组件部署

本环节需要完成 MySQL 服务的启动和 Hive 的安装、配置和验证。已安装 Hadoop 及需要配置前置环境。具体部署要求如下：

1、 启动 MySQL 数据库，创建 MySQL 数据库用户，用户名/密码：root/root123。把启动命令和结果截图。

2、 进入 MySQL 控制台，创建 hive 数据库，并创建 hive 用户可访问该库的所有表的所有权限，hive 用户的密码为 hive，把执行语句和结果截图。

3、 解压安装 Hive 到路径 /usr/local/hive，把执行命令和结果截图。

4、 修改/etc/profile 文件，配置 Hive 环境变量，并使之生效，将环境变量配置内容截图。

5、 把 MySQL 驱动 mysql-connector-java-5.1.26-bin.jar 复制到 hive 安装路径的 lib 目录下，把执行命令和结果截图。

6、 修改 hive-site.xml 文件，以使用上面在 MySQL 里创建的 hive 数据库保存 hive 元数据，把修改后的文件内容截图。

7、 初始化 Hive 元数据，把执行命令和结果截图。

8、 启动 hive，并验证 Hive 是否安装成功，将运行结果截图。

三、Spark 组件部署

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，安装 spark 具体部署要求如下：

1、 下载、安装并配置 spark 。

2、 配置 spark 环境变量。

3、 启动 spark shell，验证安装完的 spark 是否可用。

四、Sqoop 组件部署

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，具体部署要求如下：

1、 下载、安装并配置 Sqoop，将其安装到/usr/local/sqoop 路径下，安装完成后进行截图保存。

2、 修改 Sqoop 环境变量，并使环境变量只对当前 root 用户生效。

3、 测试 Sqoop 连接 MySQL 数据库是否成功，截图并保存结果。

任务二：数据采集（20 分）

1、分析网站，利用 chrome 查看网页源码，分析零售网站网页结构。打开零售网站（网址见附录或见资料文件夹），在网页中检查网站，浏览网站源码查看所需内容。

2、从零售网站中爬取需要数据，按照要求使用 Python 语言编写并编写爬虫代码，爬取指定数据项，有效数据项包括但不限于：卡号、商品 ID、品牌、产品名称、最小可用单位、SRP、毛重、净重、是否环保包装、是否低脂、子产品、产品类别、产品部、产品族等字段等多项字段。并将代码文件与代码截图保存。

具体步骤如下：

1）创建爬虫项目

2）构建爬虫请求

3）按要求定义相关字段

4）获取有效数据

5）将爬取到的数据保存到指定位置

3、至此已从零售网站中爬取了所需数据，下一步我们要将爬取结果进一步进行相关数据操作，请将操作命令截图并保存。

任务三：数据清洗与分析（25 分）

现已从相关网站及平台获取到原始数据集，在不涉及客户安全数据或者一些商业性敏感数据的情况、不违反系统规则条件下，对真实数据进行改造并提供测试使用。

以 product.csv 文件为例，该文件中包含了有关产品信息的数据，但原始数据经过多次采集汇总，数据集中不可避免地存在一些数据缺失、冗余、重复等现象。你的小组需要通过编写代码或脚本完成对文件 product.csv 中产品信息数据的清洗和整理，并完成数据计算和分析任务。

1、 缺失值处理

缺失值是一种常见的脏数据情况，现有数据集中某个或某些属性的值是不完全的。对于缺失值的处理，从总体上来说分为缺失值删除和缺失值插补。当缺失值过多时，信息条目本身的价值也会随之降低，此时如果对缺失值进行填补则将产生结果的人为干预。请使用 Java 语言编写 MapReduce 程序删除 product.csv 文件中缺失值（空字符串）大于 n（3）个字段的数据条目剔除原始数据集并将其输出结果文件重命名为 clean_data1.csv,并在控制台输出剔除的条目数量，截图并保存结果。

2、 重复数据处理

原始数据集来自于多个平台及网站，且为多次采集汇总，因此数据集中的某些字段有可能会出现一些重复或非法格式，例如多次采集过程中产生的重复信息，或来自于某网站的不合规数据。这些信息的存在既无实际的业务分析意义，甚至还会影响最终分析结果。请使用 Spark 程序删除 clean_data1.csv 文件中的非法数据和重复数据，将其输出至 HDFS 文件系统中，截图并保存结果。

3、 导入数据

启动 Hive。在 Hive 中创建数据库 db1_hive，在该数据库上创建表 sales、表 retail、表 product、表 custom。其表结构与 sales.csv、retail.csv、product.csv、custom.csv 相同，编写命令行查看各个表结构，将运行结果截图并保存。在 Hive

端使用命令将文件 sales.csv、retail.csv、product.csv、custom.csv 对应导入到数据库 db1_hive 的 sales 表、retail 表、product 表、custom 表中。分别验证查看数据库表总记录数量，将运行结果截图并保存。

4、 工作类型分析

在销售行业中，有这样一句话——“顾客就是上帝”。对在零售网站中注册的客户进行有效的分析，显得尤为重要。在客户数据集中，记录了客户的账号、姓名、卡片等级、地域、工作类型、孩子数量等情况。请使用 Spark 程序根据 custom.csv 文件中的数据，分析零售网站中客户的工作类型所对应的客户数量，将结果输出至 HDFS 文件系统中，将运行结果截图并保存。

5、 客户等级分析

在零售网站中，客户的等级代表了客户的购买能力，而客户的购买能力与诸多因素有关，例如年收入越多，购买能力越强；家中有小孩的客户，需要消耗更多的商品等等。请使用 Spark 程序根据 custom.csv 文件中的数据，查询零售网站中年收入在$30K - $50K 之间、在家孩子数量大于 0 的客人的信息，将结果输出至 HDFS 文件系统中，将运行结果截图并保存。

6、 媒体推广形式分析

促销是通过向市场和消费者传播信息，以促进销售、提高业绩。零售商品网站也会在不同时期，不同区域，通过不同的媒介，采用不同的促销方式进行促销活动。请使用 Spark 程序根据 custom.csv 文件中的数据，统计零售网站中不同媒体推广形式对应的总成本和总天数的情况，将结果输出至 HDFS 文件系统中，将运行结果截图并保存。

使用 Hive 系统中的 sales 表中的数据作为数据源，使用 Hive 命令，统计每种媒体推广形式的总成本、总天数，同时将数据写入数据表中，将命令与执行结果截图并保存。

7、 客户家庭信息分析

若要根据客户的特定信息了解客户在网站的消费情况，需要对客户数据及零售记录进行分析。请以 custom 表、retail 表、product 表中的数据作为数据源（custom 表中的 id 列对应 product 表中的 Product_ID 列；custom 表中的 id 列对应 retail 表中的 Customer），使用 Hive 命令，查询零售网站中卡号、卡片等级、

年收入、在家孩子数量、有车数量、产品名称、购买产品数量、总金额等信息，同时将数据写入数据表中，将语句及输出结果截图并保存。

8、 客户类型分析结果迁移

在 Hive 中创建数据库 db2_hive，并在该库中创建表 job_type_hive，包含两个字段：工作类型及客户数量，将任务三中客户类型分析结果迁移到 job_type_hive 表。在 MySQL 中创建数据库 DB，并在该库中创建表 job_type_sql，用于存储 db2_hive 中的 job_type_hive 表的数据，二者表结构相同。使用 sqoop 命令将 Hive 中的 db2_hive 库的 job_type_hive 表中数据导入到 MySQL 的 DB 库中的 job_type_sql 表，将该命令截图并保存。查看 job_type_sql 表的数据，将该命令和结果截图并保存。

9、 媒体推广形式分析结果迁移

在 Hive 中的数据库 db2_hive 创建表 media_type_hive，包含两个字段：媒体推广类型及促销数量。将任务三中媒体推广形式分析结果迁移到 media_type_hive 表。在 MySQL 中创建数据库 DB，并在该库中创建 media_type_hive 表，用于存储 db2_hive 中的 media_type_hive 表的数据，二者表结构相同。使用 sqoop 命令将 Hive 中的 db2_hive 库的 media_type_hive 表中数据导入到 MySQL 的 DB 库中的 media_type_sql 表。将该命令截图并保存。查看 media_type_sql 表的数据，查看命令和结果截图并保存。

10、 在家孩子数量与消费分析结果迁移

在 Hive 中的数据库 db2_hive 创建表 child_num_hive，包含两个字段：在家孩子的数量和购买产品的总金额。将任务三中客户家庭信息中的在家孩子的数量和购买产品的总金额的分析结果插入到 child_num_hive 表。在 MySQL 中创建数据库 DB，并在该库中创建 child_num_sql 表，用于存储 db2_hive 中的 child_num_hive 表的数据，二者表结构相同。使用 sqoop 命令将 Hive 中的 db2_hive 库的 child_num_hive 表中数据导入到 MySQL 的 DB 库中的 child_num_sql 表。将该命令截图并保存。查看 child_num_sql 表的数据，查看命令和结果截图并保存。

任务四：数据可视化（20 分）

本任务使用数据分析统计与数据可视化终端来完成。为更好的将数据分析结

果表达出来，需要对数据分析的结束进行可视化呈现，可视化呈现，要求使用 Python 的 Django 框架或 Flask 框架编写基于 Web 的程序，程序基础框架已搭建完成，相应的数据已给出，在前端页面中，使用 Jinja2 模板引擎获取相关统计图表数据并传递给前端页面中相应的 EChart 组件。

1、 可视化的准备工作

在 windows 本机安装 Mysql 服务器及客户端，相关文件详见附录。将现有的的 retail_store_db.sql 文件导入数据库作为可视化分析的数据源。

2、 客户类型分析可视化

各种工作类型的客户数量，可以帮助商家分析哪种工作类型对应的消费者较多。根据表中数据，以柱状图呈现二者的关系。

3、 媒体推广形式分析可视化

不同媒体推广类型及其促销数量，可以帮助商家做促销活动时选择媒体推广形式。根据表中数据，以折线图呈现二者的关系。

4、 在家孩子数量与消费分析可视化

不同在家的孩子数量及其对应的购买产品的总金额，可以帮助商家分析消费者的购买行为，进而制定相应的售卖计划。根据表中数据，以直方图呈现二者的关系。

任务五：综合分析（15 分）

假定你为零售业的某店主，在综合理解任务一、二、三、四的相关结论后，对该零售业情况进行分析，并编写输出分析报告。

根据上述任务中的结论，分析以下内容：

1、 消费者中哪种工作类型的群体较多，并根据你的理解说明一下原因。

2、 简述孩子数量与家庭消费的关系，并简要分析这种现象。

3、 为了促进消费，请你制定一套详细的商品促销计划（从成本、时间、人群等方面进行阐述）。