题目：电商平台用户行为数据清洗与分析

【要求：总分 100 分，需要按照要点答题，将答题结果标注或说明。答题时间 8 小时】

一、 背景

作为一家电商平台的数据分析师，您的任务是分析用户的购物行为以提高销售额和用户满意度。平台每天收集大量的用户行为数据，包括点击、搜索、购买等活动。这些数据被用来理解用户偏好、预测趋势并优化库存管理。然而，由于网络问题、用户误操作和系统错误等原因，原始数据集中存在大量的噪音和不一致性。

二、 任务

任务 1：数据采集流程（40 分）

【要求：需要标注开始、结束时间，共 6 张表，共 40 分，每张表采集时间不超过 30 秒得满分，每张表每超过 1 秒扣 0.5 分，最多扣 10 分】

采用数据采集工具，将文本文件读入数据库，数据总量和数据表结构保持一致。

任务 2：设计清洗流程，解释每个步骤的目的和操作（60 分）

【要求：按要求完成以下 6 项任务，每项任务 10 分，无结果得 0 分，结果不准确酌情扣分】

识别并处理以下常见的数据质量问题：

1. 请识别出购买行为表的重复的记录，并将结果保存至新表，表名为”原始表名_reslut”
2. 请识别出浏览行为缺失值,并将结果保存至新表，表名为”原始表名

_reslut”

3. 请识别出搜索行为异常值（如不合理的价格或数量）,并将结果保存至新表，表名为”原始表名_reslut”
4. 请识别出评价行为格式错误 (如日期格式不一致) ,并将结果保存至新表，表名为”原始表名_reslut”
5. 请识别出至少 3 条不一致信息，如同一用户购买和评价信息不一致的数据，识别数据结果表名为“不一致要素_result”。
6. 请生成数据清洗质量报告，并使用分析图表展示清洗结果。

说明：电商平台用户行为数据的模型。这通常包括以下几个方面：

- 定义数据模型
- 用户信息（用户 ID、性别、年龄、地区等）
- 商品信息（商品 ID、类别、价格、品牌等）
- 浏览行为（浏览 ID、用户 ID、商品 ID、浏览时间、浏览的商品 ID、停留时间等）
- 搜索行为（搜索 ID、用户 ID、商品 ID、搜索时间、搜索关键词、搜索结果点击等）
- 购买行为（购买 ID、用户 ID、商品 ID、购买时间、购买的商品 ID、数量、金额等）
- 评价行为（用户 ID、购买 ID、评价时间、评价的商品 ID、评分、评论内容等）

三、 数据访问

https://docs.srdcloud.cn/file-invite/GchvVtk5TrLwjwDxBj4Kzul0P2a5/

四、 硬件资源

- 存储容量：10G。
- 计算能力：8Vcpu/16G。
- 数据采集工具：ETL（提取、转换、加载）工具，用于从不同的数据源收集数据。
- 数据库管理系统：SQL 或 NoSQL 数据库，用于存储和管理数据。
- 数据清洗开发：如 Java、Python，用于数据的清洗和预处理。
- 版本控制系统：如 Git、研发云，用于跟踪和管理数据和代码的变更历史。

五、验证方法

通过数据库，对比原始数据和清洗结果数据；使用数据清洗质量报告，判断数据清洗质量。

六、计算结果

1. 采用数据采集工具，将文本文件读入数据库，数据总量和数据表结构保持一致。
2. 请识别出购买行为表的重复的记录，并将结果保存至新表，表名为”原始表名_reslut”
3. 请识别出浏览行为缺失值,并将结果保存至新表，表名为”原始表名_reslut”
4. 请识别出搜索行为异常值（如不合理的价格或数量）,并将结果保存至新

表，表名为”原始表名_reslut“

5. 请识别出评价行为格式错误 (如日期格式不一致) ,并将结果保存至新表，表名为”原始表名_reslut“
6. 请识别出至少 3 条不一致信息，如同一用户购买和评价信息不一致的数据，识别数据结果表名为“不一致要素_result”。
7. 请说明清洗质量，并使用分析图表展示清洗结果。

七、时间限定

8 小时

八、评分点

任务 1：数据采集流程（40 分）

【要求：需要标注开始、结束时间，共 6 张表，共 40 分，每张表采集时间不超过 30 秒得满分，每张表每超过 1 秒扣 0.5 分，最多扣 10 分】

任务 2：设计清洗流程，解释每个步骤的目的和操作（60 分）

【要求：按要求完成以下 6 项任务，每项任务 10 分，无结果得 0 分，结果不准确酌情扣分】

九、 备注

数据：

- 用户信息（用户 ID、性别、年龄、地区等）
- 商品信息（商品 ID、类别、价格、品牌等）

- 浏览行为（浏览 ID、用户 ID、商品 ID、浏览时间、浏览的商品 ID、停留时间等）
- 搜索行为（搜索 ID、用户 ID、商品 ID、搜索时间、搜索关键词、搜索结果点击等）
- 购买行为（购买 ID、用户 ID、商品 ID、购买时间、购买的商品 ID、数量、金额等）
- 评价行为（用户 ID、购买 ID、评价时间、评价的商品 ID、评分、评论内容等）