京东金融信贷需求分析系统

1. 问题描述

金条是京东金融旗下的一款无抵押现金贷产品，申请人只需要在京东金条申请页面填写少量的个人信息即可申请现金贷款。在开展这类信贷业务的时候，除了要评估用户的风险之外，还需要预测用户的借款需求，只有尽可能的给有借款需求的用户分配合适的额度，才能最大限度的增加资金利用率，降低成本并增加收益 。

本题目希望参赛者通过竞赛数据中的用户基本信息、在移动端的行为数据、购物记录和历史借贷信建议信贷需求分析系统 。本题目中包含了各种维度的序列数据、品类交易数据，并希望学员利用所学的 hadoop 与 spark 大数据技术，按照要求完成系大数据系统设计、程序实现和可视化展示部分。

注意：给定的数据为业务情景数据，所有数据均已进行了采样和脱敏处理，字段取值与分布均与真实业务数据不同。

2. 数据说明

我们提供了时间为 2016-08-03 到 2016-11-30 期间，用户在移动端的行为数据、购物记录和历史借贷信息，及 11 月的总借款金额。。

数据集下载地址为：链接:https://pan.baidu.com/s/1nvOOmmt 密码:e73r

（1）文件信息

文件名	数据内容
t_user.csv	用户信息表
t_order.csv	订单信息表
t_click.csv	点击信息表
t_loan.csv	借款信息表
t_loan_sum.csv	月借款总额表

（2）数据字典

文件名	字段名	字段描述
t_user	uid	用户ID
	age	年龄段
	sex	性别
	active_date	用户激活日期
	limit	初始额度
t_order	uid	用户ID
	buy_time	购买时间
	price	价格
	qty	数量
	cate_id	品类ID
	discount	优惠金额
t_click	uid	用户ID
	click_time	点击时间
	pid	点击页面
	param	页面参数
t_loan	uid	用户ID
	loan_time	借款时间
	loan_amount	借款金额
	plannum	分期期数
t_loan_sum	uid	用户ID
	month	统计月份
	loan_sum	借款总额

3. 数据分析任务

（1）任务 1

将 t_user 用户信息到 MySQL 中，其他三个文件及，t_click，t_loan 和 t_loan_sum 导入到 HDFS 中。

（2）任务 2

利用 Sqoop 将 MySQL 中的 t_user 表导入到 HDFS 中

（3）任务 3

利用 Presto 分析产生以下结果，并通过 web 方式可视化：

- 各年龄段消费者每日购买商品总价值
- 男女消费者每日借贷金额

（4）任务 4

利用 Spark RDD 或 Spark DataFrame 分析产生以下结果：

- 借款金额超过 2000 且购买商品总价值超过借款总金额的用户 ID
- 从不买打折产品且不借款的用户 ID

（5）任务 5：流式分析

利用 spark streaming 实时分析每个页面点击次数和不同年龄段消费总金额：

- 将 t_user 存放在 mysql 中（任务 1 已做）
- 编写 Kafka produer 程序
    - 将 t_click 数据依次写入 kafka 中的 t_click 主题中，每条数据写入间隔为 10 毫秒，其中 uid 为 key，click_time+”,”+pid 为 value

    - 将 t_order 数据依次写入 kafka 中的 t_order 主题中，每条数据写入间隔为 10 毫秒，其中 uid 为 key，uid+”,” +price + “,” + discount 为 value
- 编写 spark streaming 程序，依次读取 kafka 中 t_click 主题数据，并统计：
    - 每个页面累计点击次数，并存入 redis，其中 key 为”click+<pid>”，value 为累计的次数
- 编写 spark streaming 程序，依次读取 kafka 中 t_order 主题数据，并统计：
    - 不同年龄段消费总金额，并存入 redis，其中 key 为”buy+<age>”，value 为累计的消费金额