实验二

背景

蚂蚁金服拥有上亿会员并且业务场景中每天都涉及大量的资金流入和流出，面对如此庞大的用户群，资金管理压力会非常大。在既保证资金流动性风险最小，又满足日常业务运转的情况下，精准地预测资金的流入流出情况变得尤为重要。

本次实验使用的数据主要包含用户基本信息、申购赎回记录、收益率、银行间拆借利率等多个维度，旨在通过数据分析了解资金流动规律。

数据来源：https://tianchi.aliyun.com/competition/entrance/231573/information

数据描述

用户信息表（user_profile_table）

包含约 2.8 万用户的基本信息，主要包括用户的性别、城市和星座等字段。

列名	类型	含义	示例
user_id	bigint	用户 ID	1234
Sex	bigint	用户性别（1：男，0：女）	0
City	bigint	所在城市	6081949
constellation	string	星座	射手座

用户申购赎回数据（user_balance_table）

该表记录了 2013 年 7 月 1 日至 2014 年 8 月 31 日期间申购和赎回信息、以及所有的子类目信息， 数据经过脱敏处理。脱敏之后的数据，基本保持了原数据趋势。数据主要包括用户操作时间和操作记录，其中操作记录包括申购和赎回两个部分。金额的单位是分，即 0.01 元人民币。 如果用户今日消费总量为0，即consume_amt=0，则四个字类目为空。

列名	类型	含义	示例
user_id	bigint	用户 id	1234
report_date	string	日期	20140407
tBalance	bigint	今日余额	109004
yBalance	bigint	昨日余额	97389
total_purchase_amt	bigint	今日总购买量 = 直接购买 + 收益	21876
direct_purchase_amt	bigint	今日直接购买量	21863
purchase_bal_amt	bigint	今日支付宝余额购买量	0
purchase_bank_amt	bigint	今日银行卡购买量	21863
total_redeem_amt	bigint	今日总赎回量 = 消费 + 转出	10261
consume_amt	bigint	今日消费总量	0
transfer_amt	bigint	今日转出总量	10261
tftobal_amt	bigint	今日转出到支付宝余额总量	0
tftocard_amt	bigint	今日转出到银行卡总量	10261
share_amt	bigint	今日收益	13
category1	bigint	今日类目 1 消费总额	0
category2	bigint	今日类目 2 消费总额	0
category3	bigint	今日类目 3 消费总额	0
category4	bigint	今日类目 4 消费总额	0

注 1： 上述的数据都是经过脱敏处理的，收益为重新计算得到的，计算方法按照简化后的计算方式处理，具体计算方式在下节余额宝收益计算方式中描述。

注 2： 脱敏后的数据保证了今日余额 = 昨日余额 + 今日申购 - 今日赎回，不会出现负值。

收益率表（mfd_day_share_interest）

收益表为余额宝在 14 个月内的收益率表。

列名	类型	含义	示例
mfd_date	string	日期	20140102
mfd_daily_yield	double	万份收益，即 1 万块钱的收益	1.5787
mfd_7daily_yield	double	七日年化收益率（%）	6.307

上海银行间同业拆放利率表（mfd_bank_shibor）

该表记录了 14 个月期间的银行间拆借利率（皆为年化利率）。

列名	类型	含义	示例
mfd_date	String	日期	20140102
Interest_O_N	Double	隔夜利率（%）	2.8
Interest_1_W	Double	1周利率（%）	4.25
Interest_2_W	Double	2周利率（%）	4.9
Interest_1_M	Double	1个月利率（%）	5.04
Interest_3_M	Double	3个月利率（%）	4.91
Interest_6_M	Double	6个月利率（%）	4.79
Interest_9_M	Double	9个月利率（%）	4.76
Interest_1_Y	Double	1年利率（%）	4.78

实验任务

任务一：每日资金流入流出统计

根据`user_balance_table`表中的数据，编写MapReduce程序，统计所有用户每日的资金流入与流出情况。资金流入意味着申购行为，资金流出为赎回行为 。

注：每笔交易的资金流入和流出量分别由字段`total_purchase_amt`和`total_redeem_amt`表示。请注意处理数据中的缺失值，将其视为零交易。

输出格式：

```
<日期> TAB <资金流入量>,<资金流出量>
```

例如：

```
20130701    32488348,5525022
```

任务二：星期交易量统计

基于任务一的结果，编写MapReduce程序，统计一周七天中每天的平均资金流入与流出情况，并按照资金流入量从大到小排序。

输出格式：

```
<weekday> TAB <资金流入量>,<资金流出量>
```

例如：

```
Sunday    155914552,132427205
```

任务三：用户活跃度分析

根据`user_balance_table`表中的数据，编写MapReduce程序，统计每个用户的活跃天数，并按照活跃天数降序排列。

当用户当日有直接购买（`direct_purchase_amt`字段大于0）或赎回行为（`total_redeem_amt`字段大于0）时，则该用户当天活跃。

输出格式：

```
<用户ID> TAB <活跃天数>
```

例如：

```
125    24
```

任务四：交易行为影响因素分析

用户的交易行为（如：余额宝或银行卡的购买或赎回，用户的消费情况等）受到很多因素的影响。例如：用户特性（参考用户信息表`user_profile_table`），当前利率（参考支付宝收益率表`mfd_day_share_interest`以及银行利率表`mfd_bank_shibor`）。

在上面的三个任务中，我们重点研究了`user_balance_table`表中的数据。现在，请你从其他的表中**自行选取研究对象**，通过MapReduce（或其他工具），根据统计结果（也即类似于上面三个任务的结果）阐述某一因素对用户交易行为的影响。

分析示例：

- **银行利率对申购/赎回行为的影响**：可以根据`mfd_bank_shibor`表，将银行一周利率`Interest_1_W`划分为不同的区间，统计每个区间下的**日均**资金流入和流出总量，分析利率与交易资金量之间的关系。

即使你的结论是某一因素对用户的交易行为没有显著影响，这样的结果也是完全OK的。本次实验重点关注的是使用MapReduce进行统计的过程。

提交方式

提交git仓库地址或者相关文件的zip包。实验报告应包括设计思路、运行结果和可能的改进之处等。