实验 4-2 用户行为画像分析

建议课时：60 分钟

一、实验目的

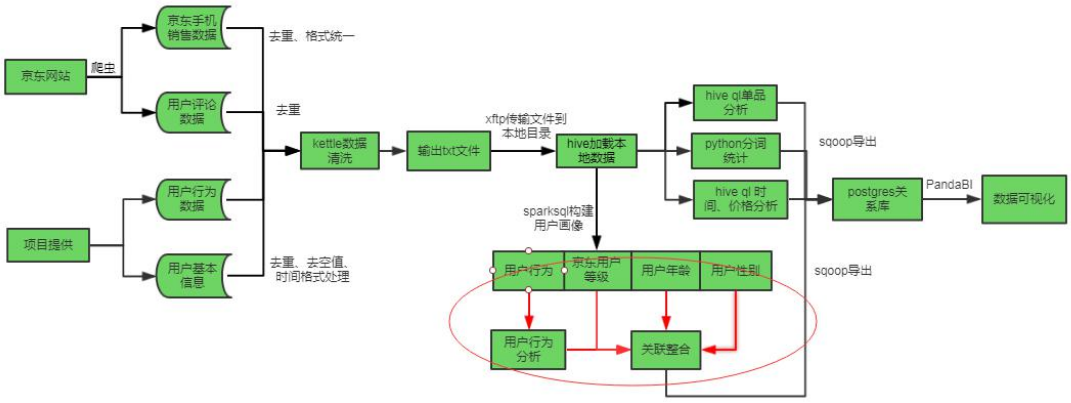
- 了解用户行为标签权重的计算方法；
- 了解用户的各标签表的关联合并；
- 了解用户行为权重标记对营销的意义；

二、实验环境

Dsight 智慧实验室的 hadoop 环境

三、实验步骤

本节实验所做内容如下红色标注：



本节实验主要是对用户行为标签表加工生成用户行为权重表，并将该表与用户属性标签表进行关联形成宽表供后续分析使用。

具体实验步骤如下：

1. 用户行为权重值的计算

1.1 行为权重介绍

用户id	用户姓名	标签id	标签名称	标签主题	行为次数	行为类型1	行为类型	行为时间	标签权重
221432	橱窗里的余甘果	A	点击	用户行为	3	0	点击	6月8日	2.34
215324	酷酷的滕	A	点击	用户行为	5	0	点击	6月12日	4.56
342126	神店通缉令	B	加入购物车	用户行为	2	1	加入购物车	8月13日	2.12
438257	撕大福	C	购买	用户行为	6	2	购买	9月12日	5.56
323422	_李默_7	D	关注	用户行为	4	3	关注	7月24日	3.46

一个用户标签表里面包括常见的字段如：用户 id、用户姓名、标签 id、标签名称、用户与该标签发生行为的次数（如搜索了两次“大数据”这个关键词）、行为类型（不同的行为类型对应用户对商品不同的意愿强度，如购买某商品>收藏某商品>关注某商品>点击某商品），行为时间（越久远的时间对用户当前的影响越小，如 5 年前你会点击某三星手机，而现在你会点击华为或苹果）。最后非常重要的一个字段是标签权重，该权重影响着对用户属性的归类，属性归类不准确，接下来基于画像对用户进行推荐、营销的准确性也就无从谈起了。

1.2 权重的计算

用户标签权重 = 行为类型权重 * 时间衰减 * TF-IDF 计算得到每个用户身上的标签权重 * 行为次数

- 行为类型权重：一般而言操作复杂度越高的行为权重越大。该权重值一般由运营人员或数据分析人员主观给出；

  自定义购买权重为 5，加入购物车权重为 4，关注权重为 3，点击权重为 2

- 时间衰减：时间衰减是指用户的行为会随着时间的过去，历史行为和当前的相关性不断减弱。

  套用牛顿冷却定律数学模型：

  F(t)=初始温度×exp(-冷却系数×间隔的时间)

  "冷却系数"是一个你自己决定的值。如果假定手机的的初始点击次数是 5 次，24 小时之后"冷却"为 1 次，那么可以计算得到"冷却系数"约等于0.067。如果你想放 慢更新率，"冷却系数"就取一个较小的值，否则就取一个较大的值。

- TF-IDF 计算标签权重：

根据 TF * IDF 即可得到该用户该标签的权重值

2. 新建用户行为权重表并加载数据

在用户行为标签表的基础上，新建用户行为权重表

用户行为标签权重表结构设计：

```
drop table if exists dwd act_weight_detail;
create table dwd act_weight_detail
(
user_id string comment '用户编码',
tag_id string comment '标签 id',
tag_name string comment '标签名称',
cnt int comment '行为次数',
tag_type_id int comment '标签类型',
act_weight_detail float comment '行为权重'
)
comment '用户画像-用户行为标签权重表';
```

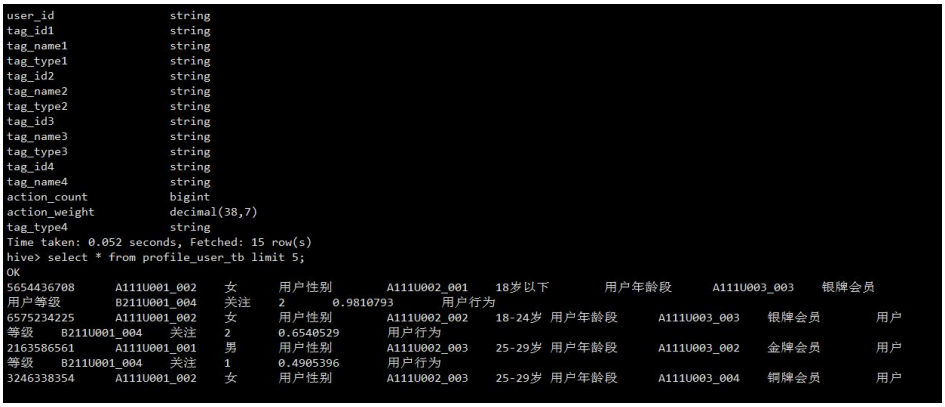
加载数据：同步用户行为标签表中的数据以及添加权重值到该表中。

3. 生成宽表

根据用户 id 关联用户所有属性标签表和用户行为权重表生成新的宽表并加载数据

表结构设计：

```
drop table if exists profile_user_tb;
create table profile_user_tb
(
user_id string comment '用户编码',
tag_id1 string comment '标签 1 ID',
tag_name1 string comment '标签 1 名称',
tag_type1 string comment '标签 1 类型',
tag_id2 string comment '标签 2 ID',
tag_name2 string comment '标签 2 名称',
tag_type2 string comment '标签 2 类型',
tag_id3 float comment '标签 3 ID',
tag_name3 string comment '标签 3 名称',
tag_type3 string comment '标签 3 类型',
tag_id4 string comment '标签 4 ID',
tag_name4 string comment '标签 4 名称',
action_count bigint comment '标签 4 行为次数',
action_weight decimal(38,7)    comment '标签 4 权重',
tag_type4 string comment '标签 4 类型'
)
comment '用户画像-用户标签宽表';
```



```
user_id                 string
tag_id1                 string
tag_name1               string
tag_type1               string
tag_id2                 string
tag_name2               string
tag_type2               string
tag_id3                 string
tag_name3               string
tag_type3               string
tag_id4                 string
tag_name4               string
action_count            bigint
action_weight           decimal(38,7)
tag_type4               string
Time taken: 0.052 seconds, Fetched: 15 row(s)
hive> select * from profile_user_tb limit 5;
OK
5654436708      A111U001_002    女      用户性别        A111U002_001    18岁以下        用户年龄段      A111U003_003    银牌会员
用户等级        B211U001_004    关注    2       0.9810793       用户行为
6575234225      A111U001_002    女      用户性别        A111U002_002    18-24岁 用户年龄段      A111U003_003    银牌会员        用户
等级    B211U001_004    关注    2       0.6540529       用户行为
2163586561      A111U001_001    男      用户性别        A111U002_003    25-29岁 用户年龄段      A111U003_002    金牌会员        用户
等级    B211U001_004    关注    1       0.4905396       用户行为
3246338354      A111U001_002    女      用户性别        A111U002_003    25-29岁 用户年龄段      A111U003_004    铜牌会员        用户
```

四、实验成果

本次实验完成后，需要得到以下结果：

- 创建用户行为标签权重表并加载数据；
- 关联合并用户属性标签表和用户行为权重表；