实验 3-2 基于 Spark 构建用户画像

建议课时：60 分钟

一、实验目的

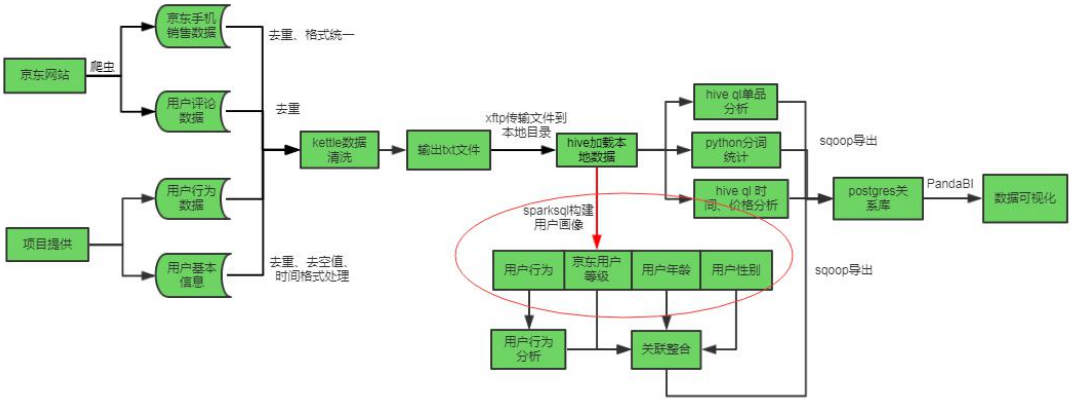
- 了解什么是用户画像；
- 熟悉用户画像的开发流程；
- 了解电商用户画像如何构建；
- 熟悉使用 Hive、Spark SQL 进行数据开发；

二、实验环境

Dsight 智慧实验室中的 hadoop 实验环境

三、实验步骤

本节实验所做内容如下红色标注：



本节实验主要是通过 sparksql 整合 hive 实现用户画像的标签开发。

具体实验步骤如下：

1. 用户画像介绍

用户画像的核心工作是为用户打标签，打标签的重要目的之一是为了让人能够理解并且方便计算机处理，如，可以做分类统计：喜欢 iphone 的用户有多少？喜欢 iphone 的人群中，男、女比例是多少？也可以做数据挖掘工作：利用聚类算法分析，喜欢 iphone 的人年龄段分布情况。

2. 构建用户画像

2.1 标签的命名

标签主题	标签类型	开发方式	是否互斥关系	用户维度
A：用户属性 B：用户行为 C：用户消费 D：风险控制	1：分类型 2：统计型	1：统计型 2：算法型	1：互斥关系 2：非互斥关系	C：cookieid U：userid

标签主题：用于刻画属于那种类型的标签，如用户属性、用户行为、用户消费、风险控 制等多种类型，可用 A、B、C、D 等字母表示各标签主题；

标签类型：标签类型可划为分类型和统计型这两种类型，其中分类型用于刻画用户属于 哪种类型，如是男是女、是否是会员、是否已流失等标签，统计型标签用于刻画统计用　户的某些行为次数，如收藏次数、近 30 日购买次数等标签，这类标签都需要对应一个　用户相应行为的权重次数；

开发方式：开发方式可分为统计型开发和算法型开发两大开发方式。其中统计型开发可 直接从数据仓库中各主题表建模加工而成，算法型开发需要对数据做机器学习的算法处　理得到相应的标签；

是否互斥标签：对应同一级类目下（如一级标签、二级标签），各标签之间的关系是否 为互斥，可将标签划分为互斥关系和非互斥关系。例如对于男、女标签就是互斥关系，　同一个用户不是被打上男性标签就是女性标签，对于高活跃、中活跃、低活跃标签也是　互斥关系；

用户维度：用于刻画该标签是打在用户唯一标识（userid）上，还是打在用户使用的设 备（cookieid）上或其他的唯一标识。可用 U、C 等字母分别标识 userid 和 cookieid 维度。

示例：对于用户是男是女这个标签，标签主题是用户属性，标签类型属于分类型，开发 方式为统计型，为互斥关系，用户维度为 userid。这样给男性用户打上“A111U001_001”， 女性用户打上标签“A111U001_002”，其中“A111U”为上面介绍的命名 方式，“001”为一 级标签的 id，后面对于用户属性维度的其他一级标签可用“002”、“003” 等方式追加 命名，“_”后面的“001”和“002”为该一级标签下的标签明细，如果是划分高、中、低活 跃用户的，对应一级标签下的明细可划分为“001”、“002”、“003”。

标签id	标签名称	标签汉语	序号	标签主题	一级标签id	一级标签
A111H001_001	male	男	1	用户属性	1	性别
A111H001_002	famale	女	2	用户属性	1	性别
A121H002_001	beijing	北京	1	用户属性	2	省份
A121H002_002	hebei	河北	2	用户属性	2	省份
A121H002_003	henan	河南	3	用户属性	2	省份
A121H002_004	anhui	安徽	4	用户属性	2	省份
A121H002_005	jiangsu	江苏	5	用户属性	2	省份
A121H002_006	zhejiang	浙江	6	用户属性	2	省份
A121H002_007	guangdong	广东	7	用户属性	2	省份
A121H002_008	fujian	福建	8	用户属性	2	省份
A121H002_009	hubei	湖北	9	用户属性	2	省份
A121H002_010	shanghai	上海	10	用户属性	2	省份

注：本案例中标签主题为用户属性和用户行为；开发方式以统计性开发为主；用户维 度使用 userid 为唯一标识。

2.2 用户基本属性的标签开发

用户属性标签：根据用户所填写的属性开发的标签和推算出来的标签（暂时不考虑）。用于了解用户的人口属性的基本情况和按不同属性维度统计。

主要数据来源：用户基本信息表

标签开发的技术工具：sparksql 整合 hive，通过 python 编写 sparksql 代码保存为 xxx.py 可执行文件，在 hadoop 环境中的 spark 组件中运行.py 文件

运行步骤：
1.启动 hive 的 metastroe 后台进程
进入 hive 安装目录 cd /opt/hive　　执行 bin/hive --service metastore
2.运行.py 文件
cd /opt/spark
执行 bin/spark-submit /data/xx/xxx.py (.py 文件的全路径)

开发实现：这里首先确定用户属性标签表的表结构，包含哪些字段，这些字段都是什么 数据类型，用户属性表创建代码如下：

```
drop table if exists dwd person_user_tag_attribute;

create table dwd person_user_tag_attribute

(

user_id string comment '用户编码',

tag_id string comment '标签 id',

tag_name string comment '标签名称',

tag_type string comment '标签类型（主题）'

)

comment '用户画像-用户属性标签表';
```

根据用户的源数据信息，可以创建以下几个用户属性标签表：

用户性别标签表

```
hive> desc profile_tag_user_gender
    > ;
OK
user_id                 string
tag_id                  string
tag_name                string
tag_type                string
Time taken: 0.059 seconds, Fetched: 4 row(s)
hive> select * from profile_tag_user_gender
    > limit 5;
OK
1000042024      A111U001_001    男      用户性别
1000824844      A111U001_001    男      用户性别
1001277790      A111U001_001    男      用户性别
1002055134      A111U001_001    男      用户性别
1002221535      A111U001_001    男      用户性别
Time taken: 0.158 seconds, Fetched: 5 row(s)
hive>
```

用户年龄段标签表

```
hive> desc profile_tag_user_age_region;
OK
user_id                 string
tag_id                  string
tag_name                string
tag_type                string
Time taken: 0.048 seconds, Fetched: 4 row(s)
hive> select * from profile_tag_user_age_region limit 5;
OK
1002719737      A111U002_001    18岁以下        用户年龄段
1002857045      A111U002_001    18岁以下        用户年龄段
1003523670      A111U002_001    18岁以下        用户年龄段
1004745095      A111U002_001    18岁以下        用户年龄段
1004853823      A111U002_001    18岁以下        用户年龄段
Time taken: 0.139 seconds, Fetched: 5 row(s)
hive>
```

用户会员标签表

```
hive> desc profile_tag_user_grade;
OK
user_id                 string
tag_id                  string
tag_name                string
tag_type                string
Time taken: 0.064 seconds, Fetched: 4 row(s)
hive> select * from profile_tag_user_grade limit 5;
OK
1003523670      A111U003_002    金牌会员        用户等级
1005473931      A111U003_002    金牌会员        用户等级
1007721027      A111U003_002    金牌会员        用户等级
1010065267      A111U003_002    金牌会员        用户等级
1027878837      A111U003_002    金牌会员        用户等级
Time taken: 0.16 seconds, Fetched: 5 row(s)
hive>
```

2.3 用户行为的标签开发

用户行为标签：是根据用户在产品上的访问行为、下单行为提取用户标签，用于定位用 户在产品上的访问情况，进而根据用户的浏览习惯、消费偏好做推荐和营销。

主要数据来源：用户行为表

注：在项目工程实践中，数据主要来源于业务类数据表、日志数据表和埋点数据表，本次案例真实数据无法获取，简单模拟了用户的基本行为数据包括（点击、加入购物　车、购买、关注商品）

标签开发的技术工具：和用户属性标签开发相同

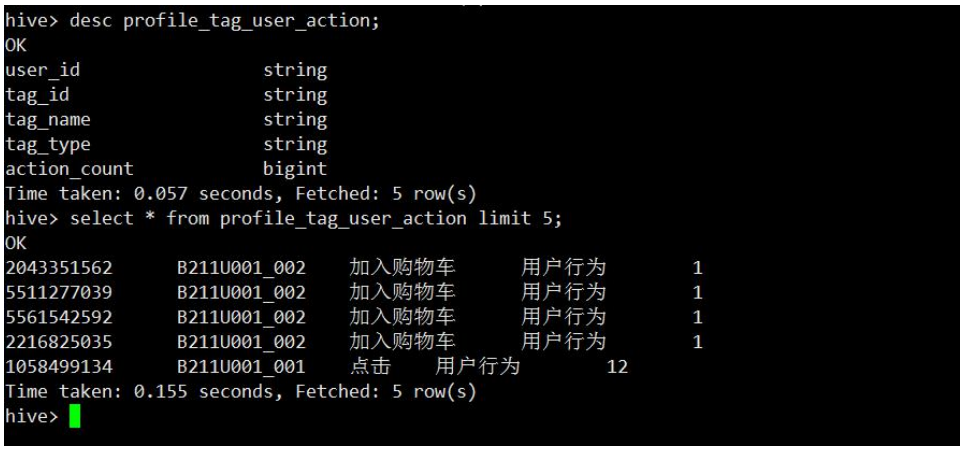
开发实现：

（1）开发用户行为标签表

```
drop table if exists dwd person_user_tag_action;

create table dwd person_user_tag_action

(

user_id string comment '用户编码',

tag_id string comment '标签 id',

tag_name string comment '标签名称',

tag_type string comment '标签类型',

action_count int comment '行为次数'

)
```

comment '用户画像-用户行为标签表';

根据用户行为表信息，通过 sql 语句创建以用户行为标签表并加载数据



四、实验成果

本次实验完成后，需要得到以下结果：

- 开发用户性别标签表
- 开发用户年龄标签表
- 开发用户等级标签表
- 开发用户行为标签表