泰迪科技数据分析项目实训

广电大数据用户画像描绘（spark 实现）
数据字段说明

培训解决方案

地址：广州市经济技术开发区开泰大道 36 号 1 栋 212

网址：http://www.tipdm.com

邮箱：services@tipdm.com

邮编：510663

联系人：

电话：

1 数据说明

业务数据表有用户基本信息表、账单信息表、订单信息表、用户状态信息变更表以及用户收视行为信息表。这些表的数据是甲方业务人员从 Elasticsearch 以 CSV 格式导出提供给乙方人员，供其进行前期预研。下面对各个表进行简要的说明。

1.1 用户基本信息表

用户基本信息表记录的是用户最新状态信息。用户基本信息表对应的 CSV 文件名称为 mediamatch_usermsg.csv，数据时间范围是 1991 年 1 月至 2018 年 6 月。用户基本信息表字段说明如表 1-1 所示。

表 1-1 用户基本信息字段说明

字段	描述
terminal_no	客户地址编号
phone_no	客户编号
sm_name	品牌名称
run_name	状态名称
sm_code	品牌编号
owner_name	客户等级名称
owner_code	客户等级
run_time	状态变更时间
addressoj	完整地址
force	宽带是否生效
open_time	开户时间

如图 1-1 所示是用户基本信息表的部分记录。

phone_no	run_time	sm_name	run_name	terminal_no	owner_name	open_time
5132880	2013-05-31 11:59:22	模拟有线电视	正常	2000417991	HC级	2013-05-31 11:59:22
5162217	2013-07-29 17:45:38	模拟有线电视	正常	2000552769	HC级	NULL
5134817	2016-06-18 10:02:11	数字电视	正常	2000156870	HC级	2012-10-29 12:21:12
5163904	2018-01-18 08:51:58	珠江宽频	欠费暂停	2000157200	HC级	2013-08-10 15:08:12
5138439	2015-11-30 23:33:04	互动电视	主动暂停	2000404015	HC级	2012-11-28 21:12:44
5143217	2013-01-19 16:35:13	互动电视	正常	2000067688	HC级	2013-01-19 16:35:13
5143998	2015-06-19 10:06:57	互动电视	欠费暂停	2000543426	HC级	NULL
5145435	2014-07-21 16:06:21	数字电视	主动销户	2000542932	HC级	NULL
5146901	2013-02-20 11:52:12	互动电视	正常	2000276061	HC级	2013-02-20 11:52:12
5164344	2013-08-19 11:38:24	互动电视	正常	2000362522	HC级	2013-08-19 11:38:24

图 1-1 用户基本信息表部分记录

1.2 用户状态信息变更表

用户状态信息变更表是用来记录用户所有时段的状态信息。用户状态信息变更表对应的 CSV 文件名称为 mediamatch_userevent.csv，数据时间范围是 1991 年 1 月至 2018 年 6 月。用户状态信息变更表的字段信息如表 1-2 所示。

表 1-2 用户状态信息变更字段说明

字段	描述
run_name	状态名称
run_time	更改状态时间
owner_code	客户等级编号
owner_name	客户等级名称
open_time	开户时间
phone_no	客户编号

如图 1-2 所示是用户状态信息变更表的部分记录。

run_name	run_time	owner_code	open_time	phone_no	owner_name
正常	2014-02-20 16:45:47	00	2014-02-20 16:45:47	3514607	HC级
欠费暂停	2016-10-11 15:07:22	00	2014-02-25 11:17:11	3514693	HC级
正常	2014-04-06 11:35:21	00	2014-04-06 11:35:21	3515712	EE级
正常	2014-02-22 12:30:24	00	2014-02-22 12:30:24	3514827	HC级
欠费暂停	2015-02-28 10:35:47	00	2014-03-08 10:49:58	3515835	HC级
正常	2014-02-23 14:39:44	00	2014-02-23 14:39:44	3514996	HC级
创建	2014-03-05 11:30:55	00	2014-03-05 18:45:23	3516482	HC级
正常	2016-07-21 10:43:13	00	2014-03-09 14:21:22	3516941	HC级
欠费暂停	2015-07-21 11:38:29	00	2014-03-08 10:49:58	3515835	HC级
欠费暂停	2015-09-16 10:20:29	00	2014-03-10 12:37:01	3517045	HC级

图 1-2 用户状态信息变更表部分记录

1.3 账单信息表

账单信息表记录用户每月的账单信息，这些账单信息会在每月一号生成。账单信息表对应的 CSV 文件名称为 mmconsume_billevents.csv，数据时间范围为 2018 年 1 月至 2018 年 7 月。账单信息表的字段如表 1-3 所示。

表 1-3 账单信息字段说明

字段	描述

fee_code	费用类型
phone_no	客户编号
owner_code	客户等级
owner_name	客户等级编号
sm_name	品牌名
year_month	账单时间
terminal_no	用户地址编号
favour_fee	优惠金额（+代表优惠，-代表额外费用）
should_pay	应收金额，单位：元

如图 1-3 所示是账单信息表的部分记录。

year_month	terminal_no	sm_name	favour_fee	owner_code	should_pay	fee_code	phone_no	owner_name
2018-03-01 00:00:00	2000304671	互动电视	0.0	00	26.5	0B	1603021	HE级
2018-06-01 00:00:00	2000304671	互动电视	0.0	00	26.5	0B	1603021	HC级
2018-04-01 00:00:00	2000016684	数字电视	5.0	00	27.0	0Y	1603120	HC级
2018-07-01 00:00:00	2000355663	数字电视	0.0	00	5.0	0Y	1603318	HC级
2018-03-01 00:00:00	2000355663	数字电视	0.0	00	5.0	0Y	1603318	HE级
2018-04-01 00:00:00	2000355663	数字电视	0.0	00	5.0	0Y	1603318	HC级
2018-01-01 00:00:00	2000355663	数字电视	0.0	NULL	5.0	0Y	1603318	HE级
2017-12-01 00:00:00	2000355663	数字电视	0.0	NULL	5.0	0Y	1603318	HE级
2018-05-01 00:00:00	2000355137	数字电视	0.0	00	5.0	0Y	1603354	HC级
2017-12-01 00:00:00	2000355137	数字电视	0.0	NULL	5.0	0Y	1603354	HE级

图 1-3 账单信息表部分记录

1.4 订单信息表

订单信息表记录用户的订购产品的信息，用户每订购一个产品，都会有相应的记录。订单信息表对应的 CSV 文件名称为 order_index.csv，数据时间范围为 2010 年 1 月至 2018 年 5 月。订单信息表的字段说明如表 1-4 所示。

表 1-4 订单信息字段说明

字段	描述
phone_no	用户编号
owner_name	客户等级名称
optdate	产品订购状态更新时间
Prodname	订购产品名称
sm_name	用户品牌名称

offerid	订购套餐编号
offername	订购套餐名称
business_name	订购业务状态
owner_code	客户等级
prodprcid	订购产品名称（带价格）的编号
prodprcname	订购产品名称（带价格）
effdate	产品生效时间
expdate	产品失效时间
orderdate	产品订购时间
cost	订购产品价格
mode_time	产品标识，辅助标识电视主、附销售品
prodstatus	订购产品状态
run_name	状态名
orderno	订单编号

如表 1-4 是订单信息表的部分记录。

offerid	offername	owner_name	orderdate	prodname	expdate	business_name
GZ122216	互动标准包(副卡)	HC级	2015-01-30 09:44:21	标清直播基本包_广州	2050-01-01 00:00:00	欠费暂停状态
GZ122216	互动标准包(副卡)	HE级	2015-01-30 09:44:21	所有基本节目_时移	2050-01-01 00:00:00	欠费暂停状态
GZ122216	互动标准包(副卡)	HE级	2015-01-30 09:44:21	个人用户免费专区_点播	2050-01-01 00:00:00	正常状态
GZ122216	互动标准包(副卡)	HE级	2015-01-30 09:44:21	基本组_点播	2015-03-31 00:00:00	到期暂停状态
00118041	[互动]优惠购机(388元)(26.5元/月)	HE级	2013-12-02 15:10:03	个人用户免费专区_点播	2014-12-31 23:59:59	到期暂停状态
GZ101369	支持单片点播权限(按片付费)	HE级	2013-12-02 15:10:03	广州基本点播组	2050-01-01 00:00:00	正常状态
GZ122216	互动标准包(副卡)	HE级	2015-01-30 09:44:21	标清直播基本包_广州	2050-01-01 00:00:00	正常状态
GZ122560	互动+联合宽带-59元包	HE级	2016-01-19 11:07:02	精彩点_点播	2050-01-01 00:00:00	正常状态
GZ122560	互动+联合宽带-59元包	HE级	2016-01-19 11:07:02	宝贝家	2050-01-01 00:00:00	正常状态
GZ122560	互动+联合宽带-59元包	HE级	2016-01-19 11:07:02	应用类点播组	2050-01-01 00:00:00	正常状态

图 1-4 订单信息表部分记录

1.5 用户收视行为信息表

用户收视行为信息表记录了用户观看电视的收视信息，其中观看方式可分为直播、点播和回看，用户每切换一个频道都会生成一条新的记录。用户收视行为信息表对应的 CSV 文件名称是 media_index.csv，数据时间范围是 2018 年 5 月至 2018 年 7 月。用户收视行为信息表的字段说明如表 1-5 所示。

表 1-5 用户收视行为字段说明

字段名	描述

terminal_no	用户地址编号
phone_no	用户编号
duration	观看时长，单位：毫秒
station_name	直播频道名称
origin_time	观看行为开始时间
end_time	观看行为结束时间
owner_code	客户等级
owner_name	客户等级名称
vod_cat_tags	vod 节目包相关信息（nested object）按不同的节目包目录组织
resolution	点播节目的清晰度
audio_lang	点播节目的语言类别
region	节目地区信息
res_name	设备名称
res_type	媒体节目类型 0 是直播，1 是点播或回看
vod_title	vod 节目名称
category_name	节目所属分类
program_title	直播节目名称
sm_name	用户品牌名称

如表 1-5 所示是用户收视行为信息表的部分记录。

terminal_no	phone_no	duration	station_name	origin_time	end_time	owner_code	owner_name
2000148366	5179844	434000	中央1台-高清	2018-10-13 22:18:21	2018-10-13 22:25:35	00	HC级
2000256434	1645503	660000	广东体育-高清	2018-10-13 21:19:00	2018-10-13 21:30:00	NULL	HE级
2000228389	1658031	3360000	广东体育-高清	2018-10-13 22:31:00	2018-10-13 23:27:00	00	HC级
2400214315	1709427	57000	湖北卫视-高清	2018-10-13 22:32:31	2018-10-13 22:33:28	00	HC级
2000212569	1405629	265000	上海纪实-高清	2018-10-13 22:22:57	2018-10-13 22:27:22	NULL	HE级
1200256367	1028894	1473000	广东体育-高清	2018-10-13 20:10:57	2018-10-13 20:35:30	00	HC级
12000499	1571196	1800000	CGTN	2018-10-13 23:30:00	2018-10-13 00:00:00	00	EE级
1200049788	1565737	76000	北京卫视-高清	2018-10-13 21:13:11	2018-10-13 21:14:27	00	HC级
1300173753	1993722	154000	北京纪实-高清	2018-10-13 23:44:38	2018-10-13 23:47:12	00	HC级
1100047425	2122375	3060000	湖北卫视-高清	2018-10-13 20:33:00	2018-10-13 21:24:00	00	HC级

图 1-5 用户收视行为表部分记录

把以上的 CSV 文件导入到 Hive 的 user_profile 库中（导入代码详情在代码清单的 csv2hive.hql。在导入 CSV 文件前，要将数据表的表头去掉，代码在 delete_head.sh）。在 Hive 的 user_profile 库中，这 5 个表分别为 mediamatch_usermsg（用户基本信息表）、

mediamatch_userevent（用户状态信息变更表）、mmconsume_billevents（账单信息表）、order_index_v3（订单信息表）、media_index_3m（用户收视行为信息表）。