一、用户画像目的

1.精准触达：特定标签组人群短信、push、电话触达。

2.统计分析：某一个学校/某一类专业学生报名付费兼职次数/人数。

3.数据挖掘：用户画像作为推荐的数据准备；利用协同过滤，为用户推荐付费兼职；利用聚类算法分析用户偏好。

4.产运评估：支持运营策略，产品改版A/B测试。

5.B端服务：针对B端定制化服务，RPO产品等。

二、青团社用户画像构建

1.技术流程

1.1.源数据搜集（后端/数据开发）

1.2.数据预处理（后端/数据开发）

1.3.行为建模（数据分析）

1.4.构建画像（数据分析）

2.源数据处理与分析

2.1.用户数据

2.1.1.用户静态数据

用户的某些属性固定不变（性别），或在长周期内（通常指一年以上）不会发生变化（年龄），或者其属性不依赖用户行为，上述用户数据统称为用户静态数据。

用户静态数据来源：

1）个人资料填写：

如性别，年龄，学校，专业，年级，学历，常住地，学校所在地。

2）清洗和补充：

数据清洗法：如性别缺失，当我们检测到用户报名过仅女性类岗位时候，补足性别；

模型清洗法：高精度聚类出现存平台女性用户报名特征，反向推出该特征向量维度下可能性别为女的概率。

2.1.2.用户动态数据

用户动态数据与用户行为密切相关，目前我们能拿到的准确行为数据更多是用户的业务数据，故我们目前初版计划先完成用户业务数据积累与分析。

主要行为动作（分端：app，小程序）：

下载，注册，登入，浏览，点击，报名，收藏，评价，搜索（商家页面，学生评价商家），

其中点击，浏览，报名等为用户重要使用数据，结合上述行为的时间序列，我们可以得到时间交叉维度的指标，后续用户画像建模中详细描述。

3.用户标签

用户标签由两块内容构成：标签名称和标签权重

e.g.

标签名称	标签权重
付费兼职爱好者	0.75
活动兼职爱好者	0.6

4.用户画像建模

4.1.用户基础属性表

根据用户所填写的属性标签和推算出来的标签。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。

性别：作为关键属性，性别无法判断的注册用户数占比达32.5%（性别缺失情况较为严重），作为完善数据的第一步，我们将首先补全未知，all，other类目下的用户性别。采用规则法和模型识别

1）规则法：当用户信息未知，通过一定规则补全；

2）产品信息不全@石头

3）信息搜集，促活小任务，问答兼职形式@鹏飞

4）模型识别：聚类算法

性别	人数
未知	1567114
ALL	1779
FEMALE	1944043
MALE	1365654
OTHER	26439

年龄：

1）产品端提示补全@石头

2）身份证信息截取

3）模型识别：暂不考虑

4）信息搜集，问答兼职回答@鹏飞

城市：

在搜集运营需求后，本周与前后端，客户端简单沟通，目前用户城市信息方案补全例子如下：

data_time	create_time	device_id	user_id	经纬度获取到城市	ip地址解析对应城市	首页用户选择城市	用户注册填写城市	端口	渠道	预留字段
2019/1/2 0:00:00	2019/1/2 0:00:00	aaaaaa	111111	87	92	107	87	ios	1	

data_time：数据写入数据库时间

create_time：数据生成时间

经纬度获取到的城市：按天搜集，搜集用户当天第一次获取到的定位

ip地址解析对应城市：按天搜集，搜集用户当天第一次session获取到的IP

首页用户选择城市：按天搜集，搜集用户当天session时左上角的位置，后续在一个session中，如果用户没有进行手动地址切换，则记录当前城市，当用户进行切换时候，重新记录整行信息

用户注册填写城市：用户注册时，或者后续补全，不做任何默认填写

用户的城市信息补全工作，在与德叔确认后，将在年后以项目形式开展

剩余用户基础信息字段：

生日、星座、城市等级、手机前几位、手机运营商、邮件运营商；

学校，学校所在地，专业，年级，学历；

上述字段将也从两个角度出发：数据清洗得到数据和补全

用户数据挖掘与算法：

1）用户忠诚度信息表（模型）

用户app，小程序忠诚度：e.g. 通过一定规则判断（用户登入次数，最近一次登入时间间隔）+ 聚类算法实现忠诚度分层：

忠诚用户，变心用户，浏览用户，一夜情用户，未识别用户；

2）用户马甲信息表

对应关系

同设备多手机号，视作同一用户特征

同一mac地址多手机号

4.2.报名单属性表

4.2.1 用户报名付费兼职属性表

该表位于大数据平台查询项目中：qtshe*query*part*job*apply*properties*by_day

以下字段构成：

用户 id,

首次报名付费兼职的时间',

首单 id',

末次报名付费时间

末单 id',

首次报名距今时间（存小时）',

末次报名距今时间（存小时）',

作为新用户报名次数',

总的报名次数',

线下兼职报名最多一级类目id',

线下兼职报名最多一级类目次数',

线下兼职报名次多一级类目id',

线下兼职报名次多一级类目次数',

线下兼职报名排名第三的一级类目id',

线下兼职报名排名第三的一级类目次数',

线下兼职报名最多二级类目id',

线下兼职报名最多二级类目次数',

线下兼职报名次多二级类目id',

线下兼职报名次多二级类目次数',

线下兼职报名排名第三的二级类目id',

线下兼职报名排名第三的二级类目次数',

线下兼职报名最多城市id',

线下兼职报名最多城市次数',

线下兼职报名次多城市id',

线下兼职报名次多城市次数',

线下兼职报名排名第三的城市id',

线下兼职报名排名第三的城市次数',

线上兼职 id',

线上兼职报名次数',

线上兼职报名最多二级类目id',

线上兼职报名最多二级类目次数',

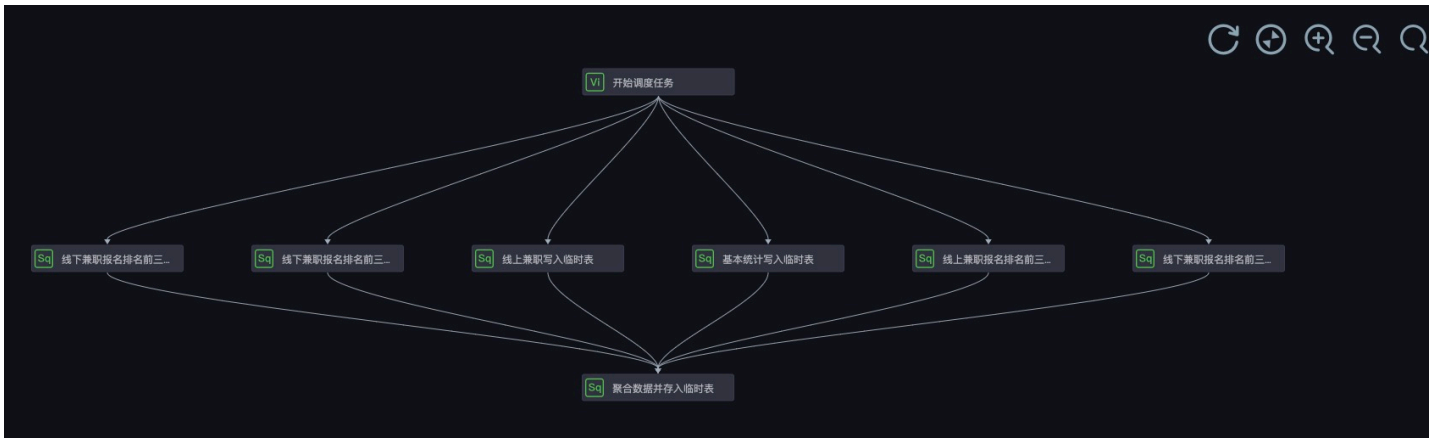
线上兼职报名次多二级类目id',

线上兼职报名次多二级类目次数',

线上兼职报名排名第三的二级类目id',

线上兼职报名排名第三的二级类目次数'

数据报表生产工作流如下：



4.2.2 用户报名付费小任务属性表

用户id,

首次领取付费小任务时间，：区别付费 非 付费 在task*apply 里面 company*id not in (165187,17101) 就是付费

首单 id',

末次领取付费小任务时间

末单 id',

首次领取距今时间（存小时）',

末次领取距今时间（存小时）',

作为新用户领取次数', account center

总的领取次数', task*center.task*apply count(*)

总的审核通过次数, task*center.task*apply count(*) where status = 100

领取各个classify*parents*id次数：目前classify*parents*id就六个，所以直接按照id分6个类目即可, task*apply classify*parent_id

审核通过各个classify*parents*id次数：目前classify*parents*id就六个，所以直接按照id分6个类目即可,

task*apply classify*parent_id where status = 100

累计获得现金，任务单审核通过 sum(price) status = 100 pay*type = 0 pay*status = 2

累计领取现金形式小任务次数，count(*) status = 100 pay*type = 0 pay*status = 2

累计获得青豆，任务单审核通过 sum(price) status = 100 pay*type = 1 pay*status = 2

累计领取青豆形式小任务次数 count(*) status = 100 pay*type = 1 pay*status = 2

分classify*parents*id累计获得青豆

分classify*parents*id累计获得现金

数据报表生产工作流如下：