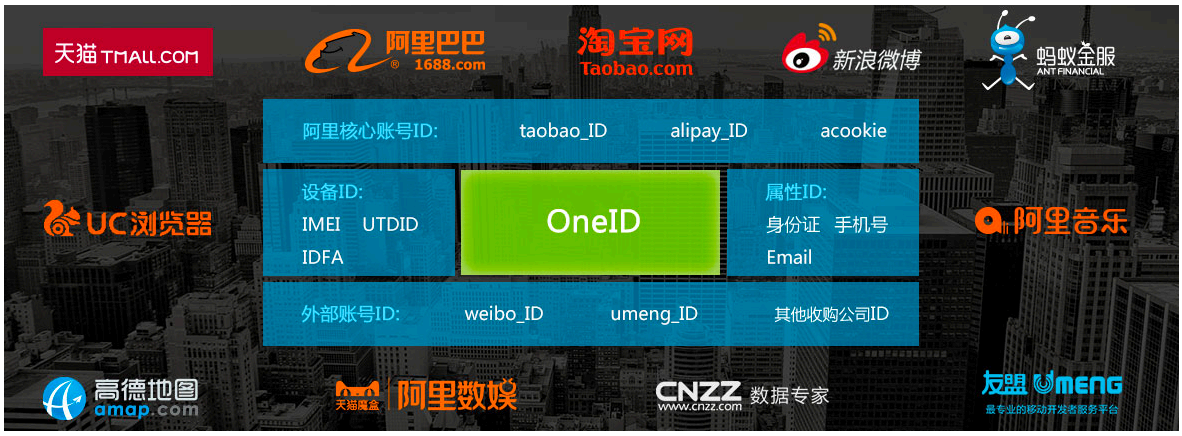
IDMapping在离线一体化解决方案

背景：



对于几乎所有的互联网企业，对于识别、追踪用户身份都有强烈的需求，精准识别用户身份后，就可以收集用户个性化的行为、身份数据，比如用户浏览了什么商品，看了什么视频，去了哪个餐厅等等，从而可以对搜索，广告，推荐等等场景做出精准、个性化的展现。

IDMapping技术路线

1、业务调研

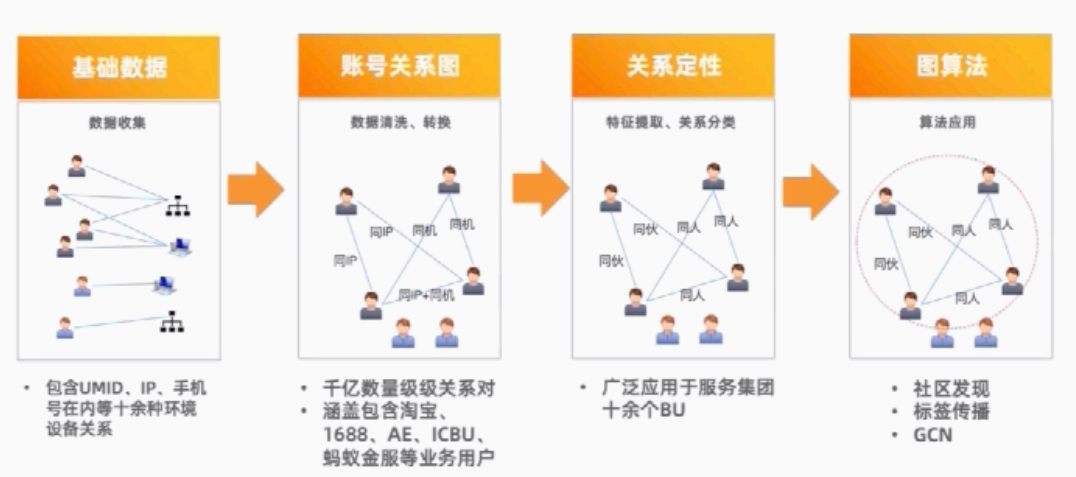
适用的业务体系，以阿里集团为例，涉及到多个场景需要使用IDMapping

跨域账号打通：

阿里集团几十个APP 账号打通，进行联合建模或者空间探查，定位到同一个自然人。为后续的营销推荐提供最基础和核心的人员定位。

同人识别：营销风控、风控策略

阿里云新人权益、大促黄牛账号拉新判定



设备账号转换：

手淘场景–站外投放：用户增长按流量来源可分为一方、二方和三方：三方是指在站外媒体投放广告，将外部媒体的流量资源引入手淘；二方是指在集团二方app（如优酷、高德等）上做广告投放；而一方是指在手淘内部做用户转化，如发push、促交易等。

2、数据梳理

跟进业务场景和业务数据情况进行整体的梳理，首先需要确定核心数据：用户ID和设备ID、用户与设备关系数据。业务方根据当前埋点数据进行抽取和盘点；

在梳理的过程中可能遇到的数据问题：数据不完整、弱关系，没有直接关系的数据，是否通过多种数据关系进行关联和挖掘。

对数据源进行细致梳理过程，可将数据源的表名、数据描述、抽取原则、数据量级、ID选取、特征选取等细节信息汇总起来，待数仓架构时作为设计依据；

完成数源梳理后，大体解决了数据从哪里来的问题，该如何抽取的问题，接下来需要进行数仓架构、数据流程和初步模型的设计工作；

3、数仓架构建设

主要涉及数据接入层和清洗层，需要将原始表信息接入并指定多种规则进行清洗。

基础层：

主要做ID关系收集、特征收集、数据抽取、粗清洗等ETL处理；基于阿里云大数据平台产品，底层大数据平台MaxCompute

准备层：

该层会将每天增量用户ID累计起来，其中包括单节点和两两ID关系对的累计，并且对各ID精清洗打标，识别ID格式异常的情况，最后还会针对用户质量特征模型，提取或汇总基础特征；

关系识别层：

通用ID（imei、idfa、tid等）兑换其他APP账号等根据两两节点的边出现的日期、pv等信息，计算关系对的活跃度；

ID构建层：

该层是最终提供用户使用的ID-Mapping结果表，主要包括通用ID、节点ID互相兑换表和业务通用兑换表等，为了方便下游常用的ID查询和join操作，在结果表提供高性能的在线查询能力，这样一来能够使下游使用更高效；

4、数据策略

基础层

该层重点工作在于数据抽取、清洗和融合等ETL处理，虽然比较基础，但该层构建是否精准，将决定了项目的成败；数据抽取在Step2中已经提及，重点是解决抽取哪些数据的问题；如：**ID清洗、ID修复以及一对多的典型场景**；

- **ID清洗：**ID清洗务必保证以下两个细节点需要特别注意，以免造成重大影响：

  【准确性】在ID清洗过程时，需要反复测试，保证清洗工作的质量，不正确的ID清洗会影响ID收容完整性和ID标识的准确性，对下游影响非常大；

  【一致性】把ID清洗封装成UDF函数，保证ID清洗的一致性，并且可以利用工厂模式实现，提高易用性与减少函数维护成本；例如：imei和idfa清洗时，使用同一个函数，传参不同而已udf_id_is_valid(imei, 'IMEI')和udf_id_is_valid(idfa, 'IDFA')；

- **ID修复：多种专家经验和策略协助**

  做一步ID修复处理，这样可以变废为宝，减少脏数据的比率例如：去除imei字符串尾部多余；（如：A1000037A357E7:中国电信:null 、869421015444548:ChinaUnicom:+8615639491616）

准备层：

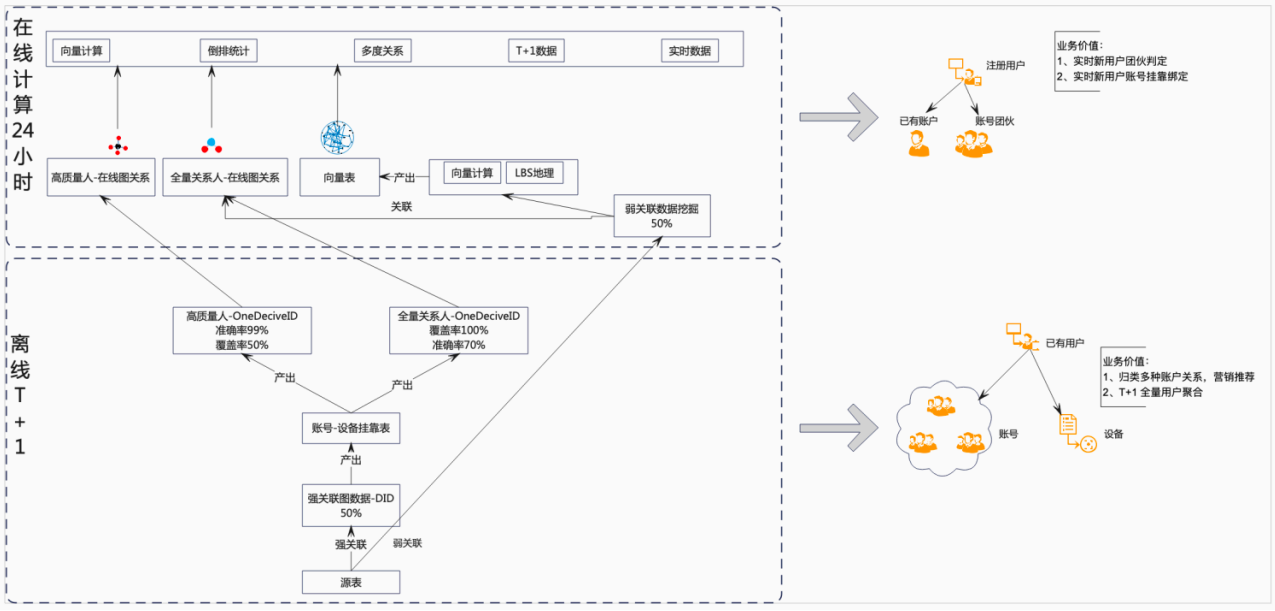
经过基础层初步处理后，我们得到了相对干净且存储量较小的轻度汇总的天级增量数据，接下来就需要将每天增量的用户ID、ID关系对累积起来，并且汇总出一些基础特征；例如：用户ID对应的设备列表、关系对出现的日期和频次、数据源分布等；另一方面，使用ID清洗统一函数进行精清洗（合法性标识），放准备层处理的主要原因是该层大多都是天级全量数据（累积汇总数据），在计算消耗方面会比在未汇总ID的基础层更加节省资源；

识别层

基于进行ID关系1对多的活跃度计算，主要进行相应的特征提取和关系分类。将这个部分能力进行输出，可广泛应用于服务各种业务BU。

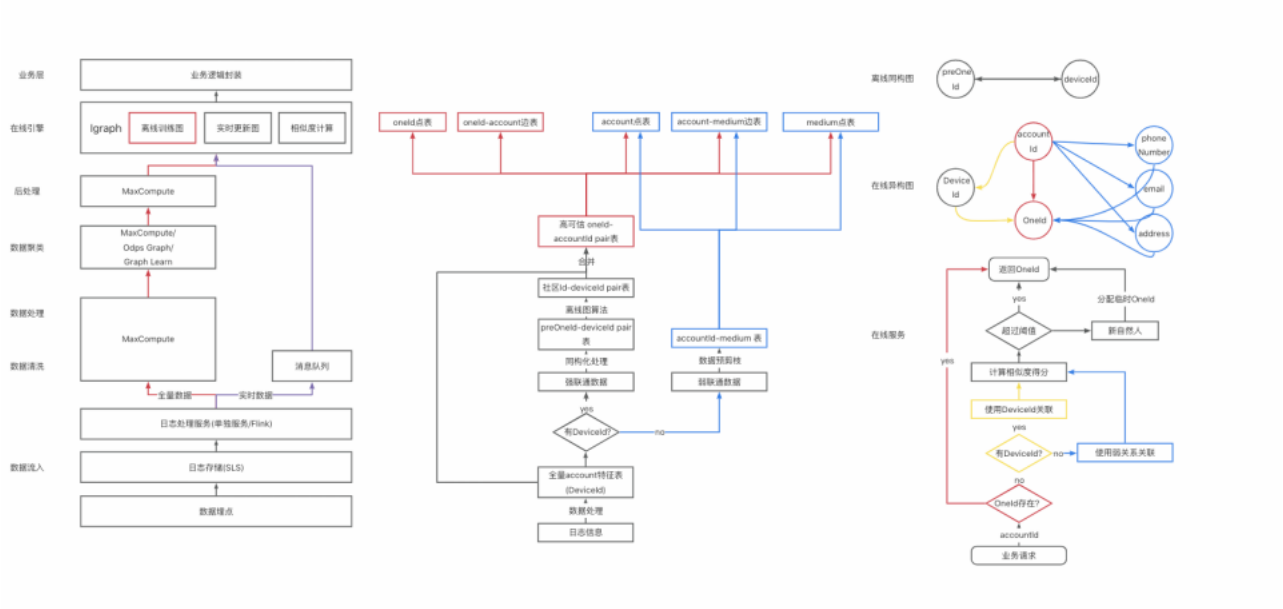
在离线IDMapping产品模块和技术组件

在离线IDMapping解决方案



技术实现细节：

https://yuque.antfin.com/igraph/xwgh5l/rbb7mk



- 数据清洗
  - 粗清洗：校验id合法性，非法id会被打标并保留，可供下游选择性使用
  - 精洗：需要反复测试以保证准确性和一致性
  - ID修复：对于部分可修复数据进行ID修复，变废为宝，减少脏数据比例
- 图算法
  - 社区发现
    - louvain：可以指定seed以提高高置信度的设备的权重，减少它们被合并的概率
    - Weakly Connected Components：通过为关联边增加权重值（活跃程度、置信程度、最近登陆等）增加划分的准确率，协助区分设备更换、账号公用等场景
  - 相似度计算
    - 在没有强关联信息的情况下，使用弱关联媒介计算账户之间的相似度，评估相同自然人使用的可能性

方案技术优势：

1、在离线一体化方案

将业界传统的T+1离线方案直接升级为24小时实时IDMapping关联能力。针对实时判定场景，提供的一体化数据生成、产出到查询链路。

同时，可提供专家经验，配合业务将IDMapping能力从0到1的梳理和落地。

2、提供中台级数据服务能力

基于阿里云大数据平台，提供中台数据架构的服务能力。

首先，配合MaxCompute数仓，基于智能运维系统，将离线计算与在线实时计算结合落地的方案。

其次，提供多种业务数据查询能力，可提供向量、倒排、复杂图查询，让业务使用更灵活。

最后，基于离线、在线一体化方案，可配合业务自行选择，赋能多种业务，如营销推荐（T+1）、实时风控（24小时实时）。

3、更好的图算法经验：

相比SparkX的解决方案，阿里云图计算团队提供更好的图算法经验。在社区发现、联通子图基于阿里达摩院和安全团队经验输出，提供最佳的解决方案。

4、数据剪枝策略：

1. 设备和账号粗筛、2. 设备和账号pair策略、3. 二手设备时间切片

5、弱关系挖掘能力

针对没有特定强关联的数据，我们也提供一些解决方案，如多因素（规则）进行综合考虑与判定、将非确定条件中的属性进行相似的关联关系计算（向量计算、地理位置计算等手段），将各种真实数据的复杂情形做一些量化方法的转换。

6、图计算引擎的实时计算能力

1、低成本

图计算Proxy–Search多行架构让集群负载更高，提高资源利用率，节省机器资源50%；同时集群负载QPS更高1倍。

2、高性能

节点拆分、多种kkv类型，在数据构建时已经将数据进行分类、同时提供可定制的截断逻辑保证查询性能；iGraph在热点key的处理经验丰富，多级cache 能够比较好的防御这类问题，同时可以支持动态扩容等；相比开源方案，查询耗时性能RT降低100%～500%

3、秒级百万更新能力

在风控领域中，OneID– 同人防控能力是通常金融、互联网企业都需要和建设的风控规则，需要实时判定用户是否违规。该类场景需要：整体数据更新量庞大，同时对图数据的查询性能要求较高，聚焦OLTP能力；GraphCompute通过最终一致性方案能够保证单节点百万QPS更新量，生效时间在1–2s，保证风控数据的实效性到秒级，从而提升识别准确率；

4、数仓一体化对接能力

离线处理平台对接，风控安全业务都会由算法、数据团队建设完整的大数据分析，基于阿里云MaxCompute数仓，我们能够无缝对接数据源，同时支持数仓快速迭代，将数仓全量数据的迭代周期最快从T天级到小时级别。

IDMapping应用场景及业务结果

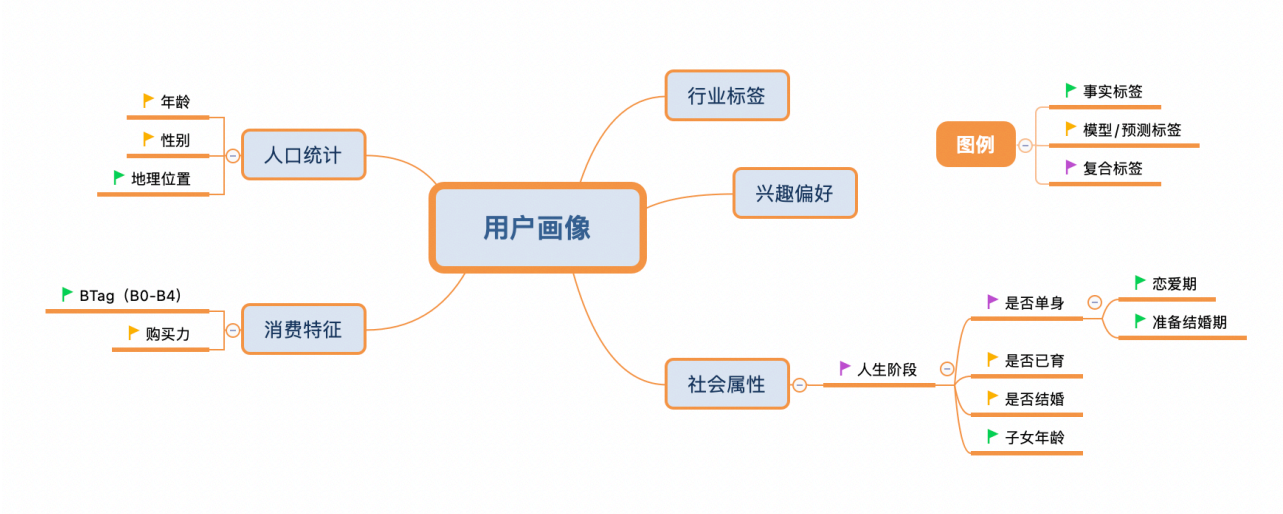
OneID 项目分为两大模块。

第一模块是IDMapping，通过技术手段将真实世界同一个人的不同类型ID识别为一个自然人，并生成稳定的互联网世界的“身份证号”，将其作为OneID的基础。

第二模块是消费者的画像系统，本模块基于IDMapping的结果对用户贴标签，形成接近用户真实画像的虚拟代表。

七猫场景的业务价值：基于IDMapping构建全域的用户画像

基于IDMapping在离线系统产出高质量用户，给这些用户做属性归类和标签能力。



在这一套高质量用户表进行相应标签能力的完善：

1）事实标签：既定事实，从原始数据中提取。例如用户设置的性别、出生日期和地理位置等。

2）统计标签：基于业务口径定义的标签，在一定周期内，业务行为的表现。例如近一个月登陆天数。

3）预测标签：使用算法产生的标签，该标签定义用户对相关业务的偏好程度，对应应该有该标签的预测的score，例如基于用户行为预测的性别、年龄等。

4）营销标签：也叫模型标签，端到端的分析模型，例如AIPL/RFM/AIDMA/AARRR模型和购买力。

使用和落地的场景：

1）分类别营销推荐

通过多渠道进行以**细分市场为中心的跨屏营销**，提升消费者的满意率、挖掘关键客户。

女性人群 – 20：言情小说、漫画

男性人群 – 20岁：武侠、科幻、玄幻。

男性人群 – 40岁：金庸武侠等

已婚夫妇：育儿类、名著、付费优质内容

家长：添加婴儿和玩具以及与儿童相关的商品。

在所有这些细分市场中，按年纪、购买力从低到高进行划分，低购买力将获得折扣，而高购买力将获得奢侈品和高质量的书籍、或者商品。

2）用户优惠判定，营销收益更高

减少对同一用户的优惠券，从而可以投资于其他买家以推动业务目标。

避免卖家创建多个帐户并使用虚假详细信息滥用补贴。

在新人优惠券场景中，需要以OneID的维度计算解决重复权益发放的问题，初步计算可以节省8%的预算，并且大量是当日注册多个账号。

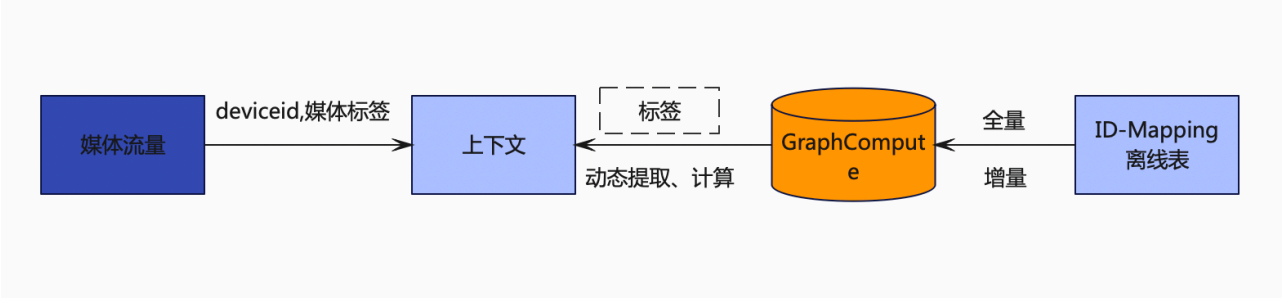
3）挖掘潜在客户，锁定客户转化

将全域IDMapping用户数据，结合用户的行为习惯做关系分析。

心动场景的业务价值：

1）精准外投：

基于IDMapping构建全域的用户画像，做精准投放，提升广告更好的转化率。转化人群的准确率，尤其是NU、RU的准确率不高，主要是由于IDMapping数据不全导致，我们可以扩展imei、appid的二级idmapping关系,例如手机号、邮箱信息等等，进行弱关系的挖掘和计算，在全域数据上提高更多的数据量，扩展IDMapping的映射量



2）实时风控，黄牛账号挖掘

识别哪些账户通过不正当手段交易或者频繁套现等复杂关系计算。在 OneID 级别制定风险评分和黑名单，以减少欺诈者重复执行欺诈案件的风险。

通过对黄牛党或者团伙行为分析，普遍的现象是最大程度的利用手上设备资源，实现账号体系的最大化；针对有部分用户通过重复注册账号领取优惠券薅羊毛的行为进行检测，需要进行用户到用户的多度查询；

根据业务的特点进行抽象定义，最终的业务逻辑可以理解为：

1）查询的业务场景：账户A – 设备G – 账号B 二跳关系

2）需要获取多种设备关联的二跳用户后，对设备路径权重加分，最终得到