网 易 数 帆 旗 下

网易易数

网易大数据用户画像实践

D I G I T A L S A I L

张长江

网易·大数据技术专家

网 易 数 帆 旗 下

网易易数

DIGITAL SAIL

目录

数据总览

网易数帆旗下

网易易数

亿级

账号日活

亿级+

可触达同人

1000+

标签沉淀

70+%

优质用户平均标签覆盖率

游戏、教育、电商、泛娱乐等

多行业产品线

参与人、流量域、位置域、关系域、

广告域、营销域、风控域等

主题解决方案

网 易 数 帆 旗 下

网易易数

网易产品线丰富，多角度描述用户

覆盖**用户娱乐、电商购物、教育、新闻资讯、通讯**等各个方面，且APP活跃度高，覆盖用户面广，积累了多维度用户行为数据，旨在**通过集团数据资产构建全域用户画像**，目前已服务于域内众多业务场景，并达成战略性合作关系

全链路数据中台产品矩阵

网易数帆旗下
网易易数

业务前台

智能分析 | 标签系统 | 用户画像 | 数字化运营 | 数字化管理 | 数字化供应链

数据中台

业务前台
- 自助查询系统
- 数据研发平台
- 数据集成工具

数据标准体系建设
- 应用数据层（ADS）
- 公共明细DWD
- 公共汇总DWS
- 数据集成工具
- 指标系统

数据治理平台
- 数据地图
- 数据学源
- 数据质量中心
- 数据资产管理

数据平台

大数据基础设施（计算和存储平台）

数据源

数据库 | 埋点采集 | 服务器日志 | 爬虫数据

网 易 数 帆 旗 下

网易易数

DIGITAL SAIL

目录

数据架构

网易数帆旗下 网易易数

整合全域数据，建设数据资产，助力集团各业务应用数据精准高效低成本（将集团数据收好，盘好，用好）

用户画像构建流程

网易数帆旗下 网易易数

依托网易大数据中台能力，搭建起完善的数仓体系，可**快速实现数据层面融合**，经多年摸索，形成**完善的用户画像研发体系**，且通过网易易数产品矩阵，为网易域内业务提供高质量的数据服务，沉淀出多行业复杂场景的各种解决方案，如增长运营、广告DMP、智能风控等

用户标签

网易大数据融合**用户娱乐、电商购物、教育、新闻资讯、通讯**等多行业10+产品线，构建起全域用户画像数据，目前**总标签数1000+**，ID量**URS、PHONE、IDFA、IMEI、OAID**等均达到**亿级**

全域用户画像建设

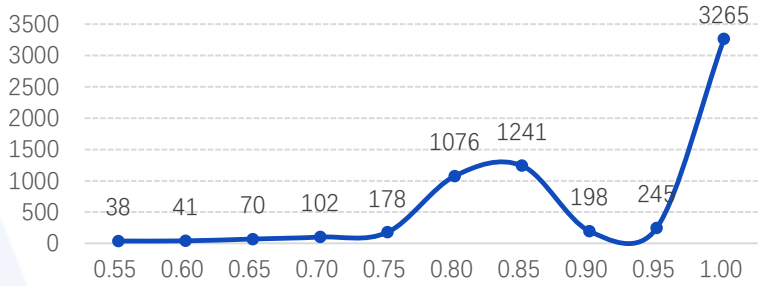
基础标签	行为标签	偏好标签	预测标签
自然属性	地域行为	出行购物	性别
教育背景	广告行为	手机数码	年龄
生活习惯	搜索行为	家装家居	近期是否出行
地理位置	全域行为	教育公益	近期是否买车
消费能力	播放行为	文化娱乐	广告内容选品
职业状况	点击行为	新闻资讯	黑产用户预测
经济情况	评论行为	金融理财	…
设备信息	关注行为	游戏竞技	
会员信息	收藏行为	动漫影视	
自定义信息	购买行为	明星艺人	
…	…	….	

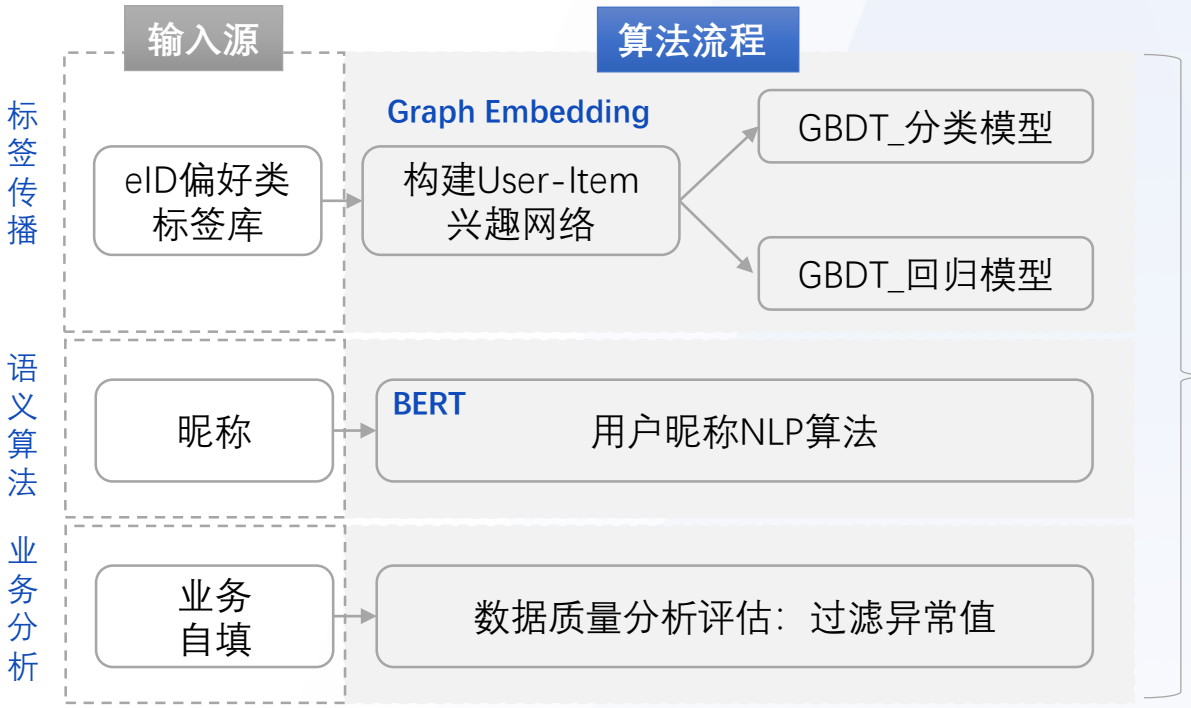
用户标签案例：性别

网易数帆旗下 网易易数

标签	原覆盖度	新覆盖度	覆盖度提升	整体准确率
性别	65%	91%	26%	89.86%

性别准确率分布（万）





模型融合：样本用户A性别

数据集	预测结果	准确率P
S1	男	0.6
S2	男	0.7
S3	男	0.8
S4	男	0.9

用户A性别：**男**，准确率 **99.21%**

IDMapping

网易数帆旗下 网易易数

1. 国家个人信息监管越来越严，用户设备ID采集越来越难，帐号打通面临巨大挑战
2. 不同业务方对设备ID类型需求不同，如果请求的设备类型与数据源中数据类型不一致，会导致无法获取数据。
3. 同一用户/设备包含不同的设备ID且ID间没有完善的映射关系，导致无法完整描述一个用户的画像。
4. 不同情境下，所谓标识用户的“账号”可能有较大的歧义，需要有所区分。

IDMapping：思路和方案

思路及方案

- 结合各种账户、各种设备型号之间的关系对，以及设备使用规律（时间和频次）等用户数据
- 采用**规则过滤**+数据挖掘**算法**（连通图划分+社区发现）判别账号是否属于同个人

可能遇到的问题	解决思路
用户可能有多个设备	使用过一定次数的设备才和账户关联
设备会过期失效（僵尸设备）	设定一个设备未使用时间衰减函数，对同时拥有多个设备的账号加大衰减力度
异常数据	需要识别出一些场景并过滤： 1. 借用朋友设备 2. 记录设备数据格式错误；有脏数据 3. 刷号等行为

识别结果示意

IDMapping流程细节：数据处理后的格式

网易数帆旗下 网易易数

ID1	ID2	参数1	参数2	参数3	参数4	参数5	.....

ID1	ID2	权值

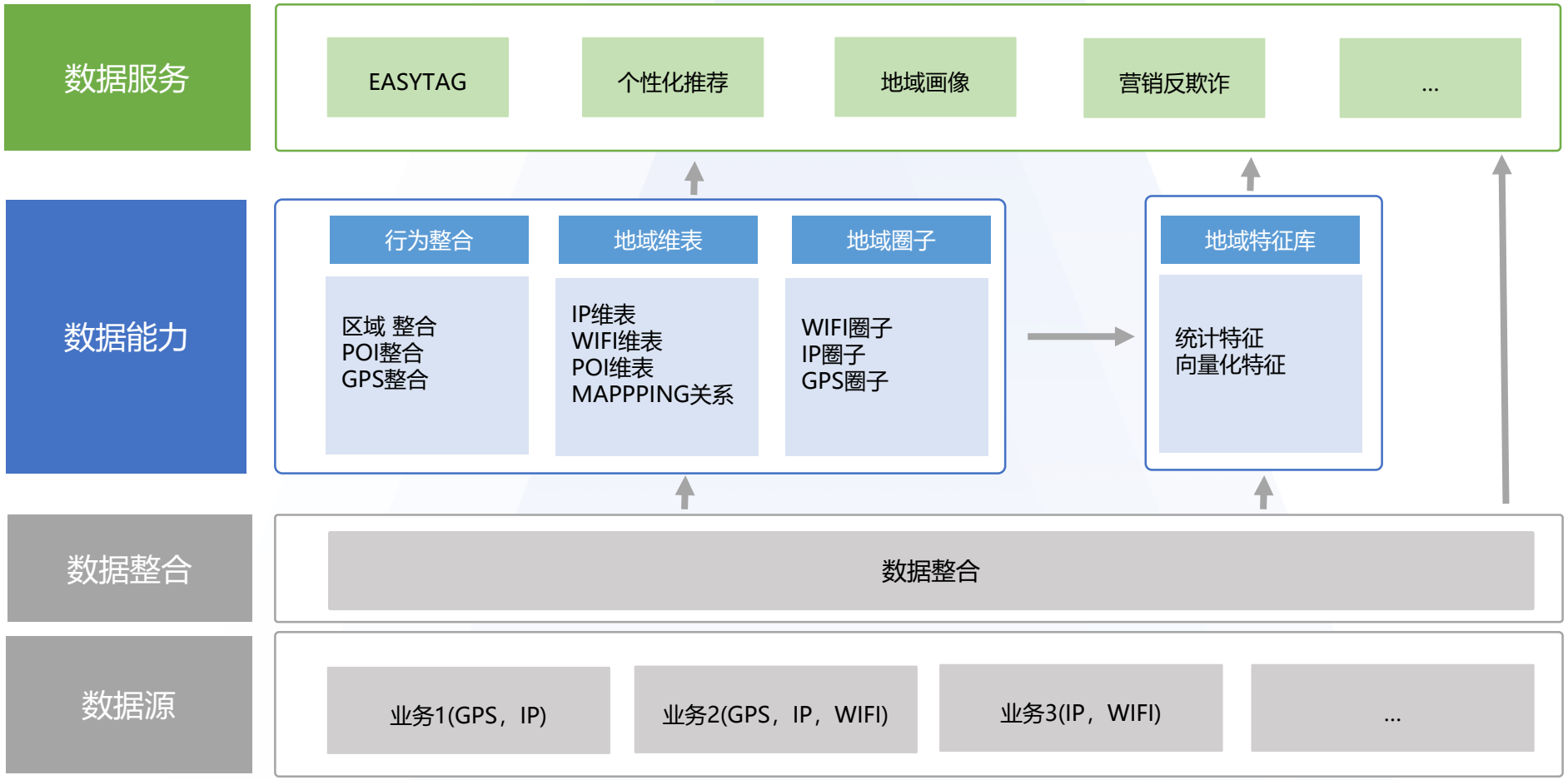
- ID1：必填
- ID2：必填
- 参数1：必填，最近采集时间，默认值（当前时间戳）
- 参数2：必填，最早采集时间
- 参数3：采集源数量
- 参数4：采集源列表
- 参数5：单位周期内出现的次数和
- 参数6：最近7天出现的次数和
- 参数7：最近30天出现的次数和

- **权值计算**：

$$权值 = \sum 参数i * 参数重要因子i$$

地域主题域

网易大数据，依托域内10+产品线，成功打造集团地域主题域数据资产，目前**IP量**级(用户量达**级)**，**HEOHASH9**级(用户量达*级)**，**POI量千万级(用户量*级)**，生产出****地域标签**



用户画像管理与存储

网易数帆旗下 网易易数

437:4#480:6

字符串拼接，使用困难

{"ent_film_news_pref7d":"4","edu_univ_news_pref7d":"6"}

大量手动命名工作，难以同步、管理

{"lv1_news_pref":{ “437":"4", “480":"6"}}

使用明文需要关联维表；扩展能力差

{"lv1_news_pref” :[{ “ID” : “437” , “CN” : “体育” , “WT” : “4” }, { “ID” : “480” , “CN” : “娱乐","WT": “6"}] }

JsonArray格式能够通用所有类型标签

网 易 数 帆 旗 下

网易易数

质量校验体系

网易包含*级真实用户数据，**每日*级**强特征用户数据，且外接*级高置信度用户数据，打造出***级真实用户属性和行为数据资产**，高可信样本集达到*级，同时具备**完善的数据校验流程以及ATB方案**

01

实名认证数据

***级身份证**认证数据

02

强特征用户数据

***级**GPS/IP数据上报

03

问卷调研+外部接入

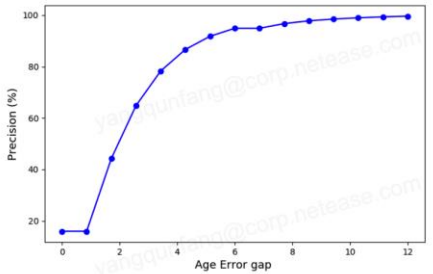
***级**外部接入高置信度用户数据

算法离线验证

mae 评估说明： mae 为2.345 相当于平均每个人预测年龄误差为2.3岁。
年龄预估的每个误差对应的准确率如下：

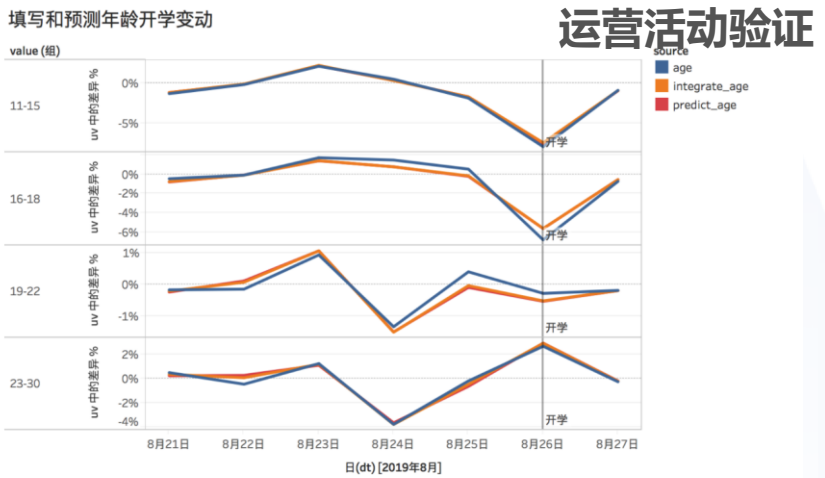
error : 0.00 precision: 15.98%
error : 2.57 precision: 64.83%
error : 3.43 precision: 78.28%
error : 4.29 precision: 86.65%
error : 12.00 precision: 99.60%

这些数据说明，预测年龄完全正确（18岁预测为18岁）的准确率为15.98%，年龄误差在4岁以内即可达到80%以上。



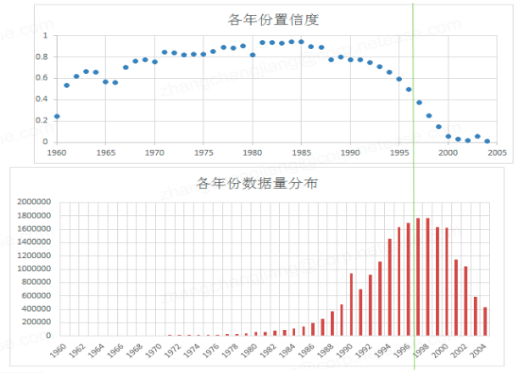
线上ABTEST

运营活动验证



真实数据验证

测试结果：置信度整体较低，平均置信度仅为0.42
原因分析：经探查各年份数量和置信度，1997年后置信度急剧下降，而数据量占总体达48.8%，从而导致整体置信度较低



质量保障体系

网易数帆旗下 网易易数

标签责任人	• **第一责任人**，快速响应业务需求和标签异常问题处理
流程优化	• 采用**端到端(End-to-End)**模式，建立快速响应机制，优化现有繁杂的开发流程，加大标签规范评审
测试&监控	• 测试：标签上线之前，**QA团队**会对标签规范和质量进行测试，并出具报告 • 预警：规范、枚举值、范围等规范，将建立**监控**机制，将问题封杀在萌芽状态
管理平台化	• 标签管理平台化（**easytag**），将标签生产、监控、应用等，工具化和产品化

网 易 数 帆 旗 下

网易易数

DIGITAL SAIL

目录

全域用户画像：应用场景

网易全域画像目前涵盖用户：**用户娱乐、电商购物、教育、新闻资讯、通讯**等多元化数据，可精准的用户定位，且覆盖用户面广，服务于**市场营销、推荐/搜索、增长运营、广告投放、智能风控**等业务目标。

市场营销

通过人群圈选、人群洞察等提升营销

增长运营

通过用户画像分析，为用户研究、用户运营等提供数据支撑

智能风控

为营销资金风险、异常用户识别提供特征或算法服务

推荐/搜索

为算法团队提供数据输入

广告投放

为广告主提供人群定向功能，提升广告投放效果

案例：用户画像案例

网易数帆旗下 网易易数

用户画像**覆盖用户****，包含基础标签、事实标签以及预测标签，覆盖用户**听歌、消费、游戏、社交、咨询**等多维度数据，对于**增长运营、个性化推荐、精准营销、智能风控**等具有重要作用

出行行为

娱乐行为

设备信息

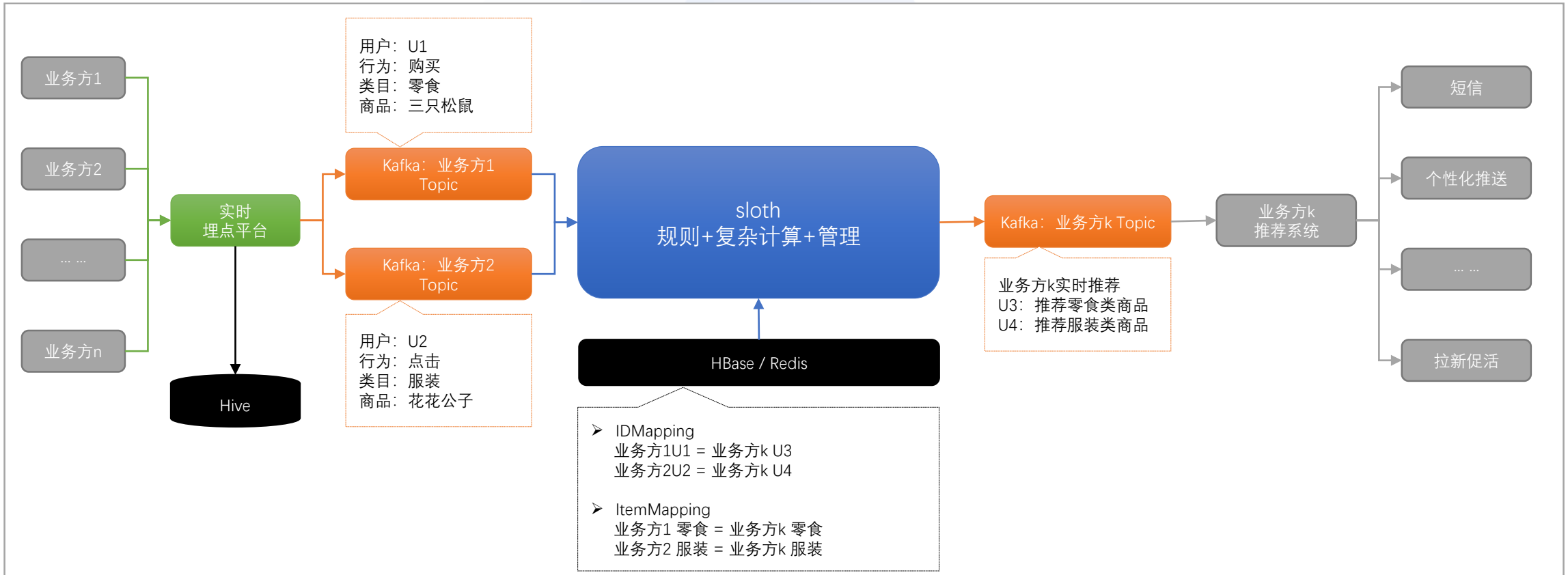
地理位置

购买行为

- 汽车：宝马X5
- 会员：黑胶年卡
- 母婴产品购买用户
- 常居一线城市
- 手机：华为P30
- 偏好歌唱主播
- 性别：女性
- 年龄：28
- 奢侈品购买用户

实时全链路推荐示例

网易集团用户数据服务实时方案，打通各业务间数据孤岛，实时融合用户数据资产，深度洞察分析用户属性，支撑各业务间数据打通和服务。



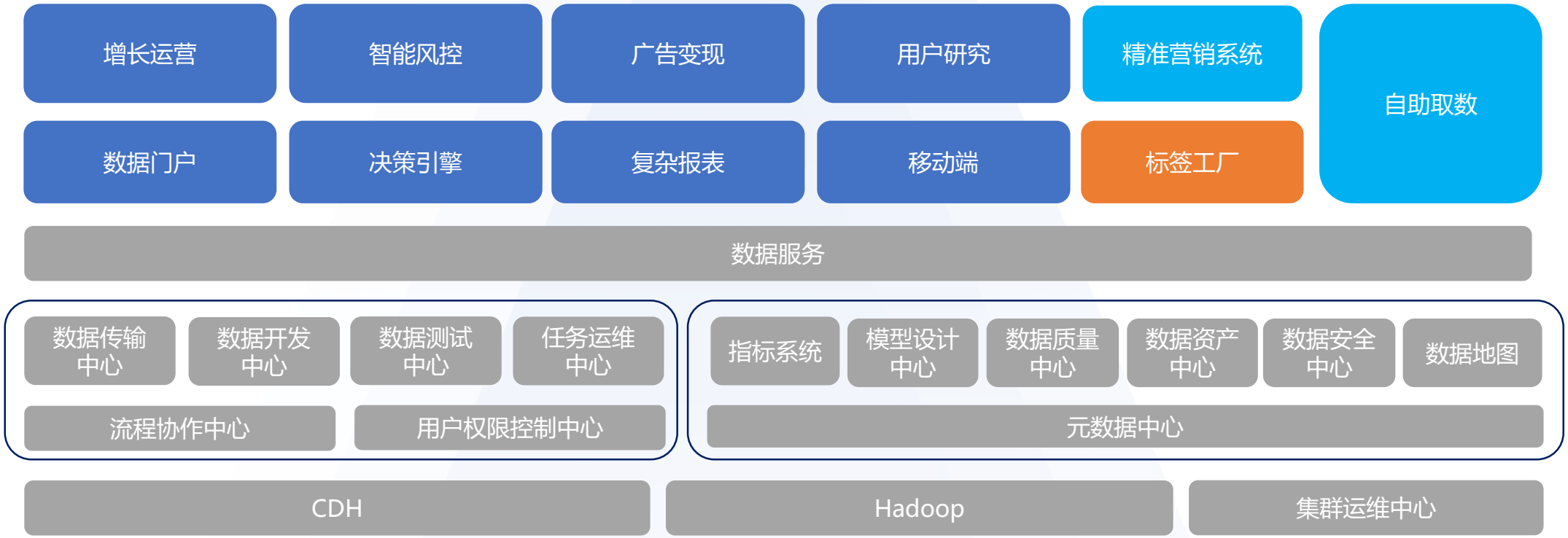
实时采集+实时计算+实时调用

数据生产力

网 易 数 帆 旗 下

网易易数

网易易数发布，从数据中台升级到数据生产力！



交流合作&招贤纳士

网 易 数 帆 旗 下

网易易数

网易易数官网

张长江

招聘官网：资深数据挖掘工程师、资深行业解决方案等
https://hr.163.com/job-list.html?currentPage=1&pageSize=10&parentProduct=P6&workPlace=229&workType=0&keyword=%E5%A4%A7%E6%95%B0%E6%8D%AE&lang=zh

THANK YOU

D I G I T A L S A I L

扫码即可关注