# 1.0达摩盘基础特征报告需求文档

## 一、项目背景

项目名称：用户画像项目达摩盘基础特征

项目在构建一个精准的用户画像标签体系，通过分析用户的行为数据，优化电商广告投放策略，提高广告效果和转化率，吸引更多商家在平台上投放广告。

项目目标：构建一个精准的用户画像标签体系，优化电商广告投放策略，提高广告效果和转化率，吸引更多商家在平台上投放广告。

**1.0**

CREATE TABLE user\_profile\_basic (

user\_id BIGINT PRIMARY KEY, -- 有一个用户ID作为主键

user\_age\_range VARCHAR(50), -- 用户年龄段区间

user\_gender VARCHAR(50)('female', 'male', 'family'), -- 用户性别倾向

user\_weight DECIMAL(5,2), -- 用户体重

user\_height DECIMAL(5,2), -- 用户身高

user\_zodiac VARCHAR(50)('Aries', 'Taurus', 'Gemini', 'Cancer', 'Leo', 'Virgo', 'Libra', 'Scorpio', 'Sagittarius', 'Capricorn', 'Aquarius', 'Pisces'), -- 用户星座

last\_updated -- 数据最后更新时间

);

user\_id：唯一的用户ID作为主键用户。

user\_age\_range、user\_gender、user\_weight、user\_height、user\_zodiac

last\_updated用于记录数据最后更新的时间，默认值为当前时间，并在每次更新时自动更新。

## 二、标签体系

**基础特征标签：**

年龄、性别、年代（此处“年代”标签在文档中未详细展开，但提及为6类标签之一，可能指更宽泛的时间分段，如“80后”、“90后”等，具体实现需进一步定义）、身高、体重、星座。

**用户年龄标签：**

定义：根据用户在淘宝平台上的行为数据，综合判断其所属的年龄段区间。

数据来源：淘宝平台用户行为日志数据、商品类目属性数据。

处理逻辑：基于出生年份直接计算或通过多行为加权判定（类目偏好、品牌偏好、价格敏感度、时间行为、搜索词分析、社交互动行为、设备信息）。

更新频率：每日更新。

**用户性别标签：**

定义：根据用户在淘宝平台上的行为数据，结合其购买商品的一级品类分布及行为频次，综合判断用户性别倾向。

数据来源：用户行为数据（浏览、收藏、加购、购买的商品记录）、商品一级品类列表。

处理逻辑：基于购买商品的一级品类分布及行为频次，判断用户性别倾向（女性用户、男性用户、家庭用户）。

更新频率：每日更新，每月全量更新。

**用户体重标签：**

定义：基于用户在淘宝系平台主动填写或设备同步的体重信息，经规则处理后生成的标准化体重标签。

数据来源：订单信息、会员资料、健康设备同步、活动表单。

处理逻辑：数据抽取、清洗转换、合并规则（优先级权重、冲突处理）、数值标准化。

更新频率：T+1日更新，设备数据实时更新。

**用户身高标签：**

定义：记录用户在淘宝、天猫等平台主动填写的实名制身高数据，以及通过商品尺码选择、客服沟通记录等渠道间接确认的身高信息。

数据来源：用户资料库、订单备注系统、客服工单系统、退货原因库、智能硬件数据。

处理逻辑：数据清洗、多源数据优先级、冲突处理规则、补全策略。

更新频率：实时更新（智能硬件数据、新订单数据）、T+1更新（用户资料修改、客服记录）、周级更新（尺码推算规则库）、月级更新（年龄身高对照表）。

**星座标签：**

定义：根据用户填写的生日信息，解析对应的十二星座。

数据来源：用户淘宝/天猫个人资料中填写的生日字段。

处理逻辑：提取月份和日期，匹配星座。

更新频率：实时更新（用户主动修改生日后）、离线批量更新（每日凌晨全量刷新）。

## 技术需求

数据存储：需要数据库（MySQL、HBase）存储用户画像数据。

数据处理：需要大数据处理框架（Hadoop）进行数据清洗、转换、合并。

实时计算：需要实时计算框架（Flink）处理智能硬件数据、新订单数据等实时数据。

消息队列：需要消息队列（Kafka）实现数据流的异步处理和解耦。

API接口：提供API接口供其他系统调用用户画像数据。

## 四、验收标准

设计文档：包括ADS层表设计、原始数据分析、关键标签实现方案、性能优化方案。

代码：实现全部标签功能，包括不分层、分层两种实现方式，代码需有明确注释。

测试文档：含测试记录及测试SQL，分层、不分层两种实现方式的数据需一致。

上线记录：包括原始数据导出、处理后的数据导出、运行日志、运行过程截图。