详解推荐系统之标签体系

机器学习与推荐算法 1月15日

嘿，记得给“机器学习与推荐算法”添加星标

来自 | 知乎

链接 | https://zhuanlan.zhihu.com/p/103129589

作者 | 龚旭东

编辑 | 机器学习与推荐算法

为什么要先介绍标签体系?

一个推荐系统效果好与坏最基本、最基础的保障是什么？如果让我来回答，一定是标签体系。我这里说的标签主要是针对物料的，对于电商平台来说就是商品；对于音乐平台来说就是每一个首歌，对于新闻资讯平台来说就是每一条新闻。与之对应的用户画像体系中那些用户实时变化的兴趣点大都也是来自于标签体系，依据用户长期和短期行为中对于物料搜索、点击、收藏、评论、转发等事件，将物料的标签传导到用户画像上，就构成了用户的实时画像和离线画像中的各个动态维度。

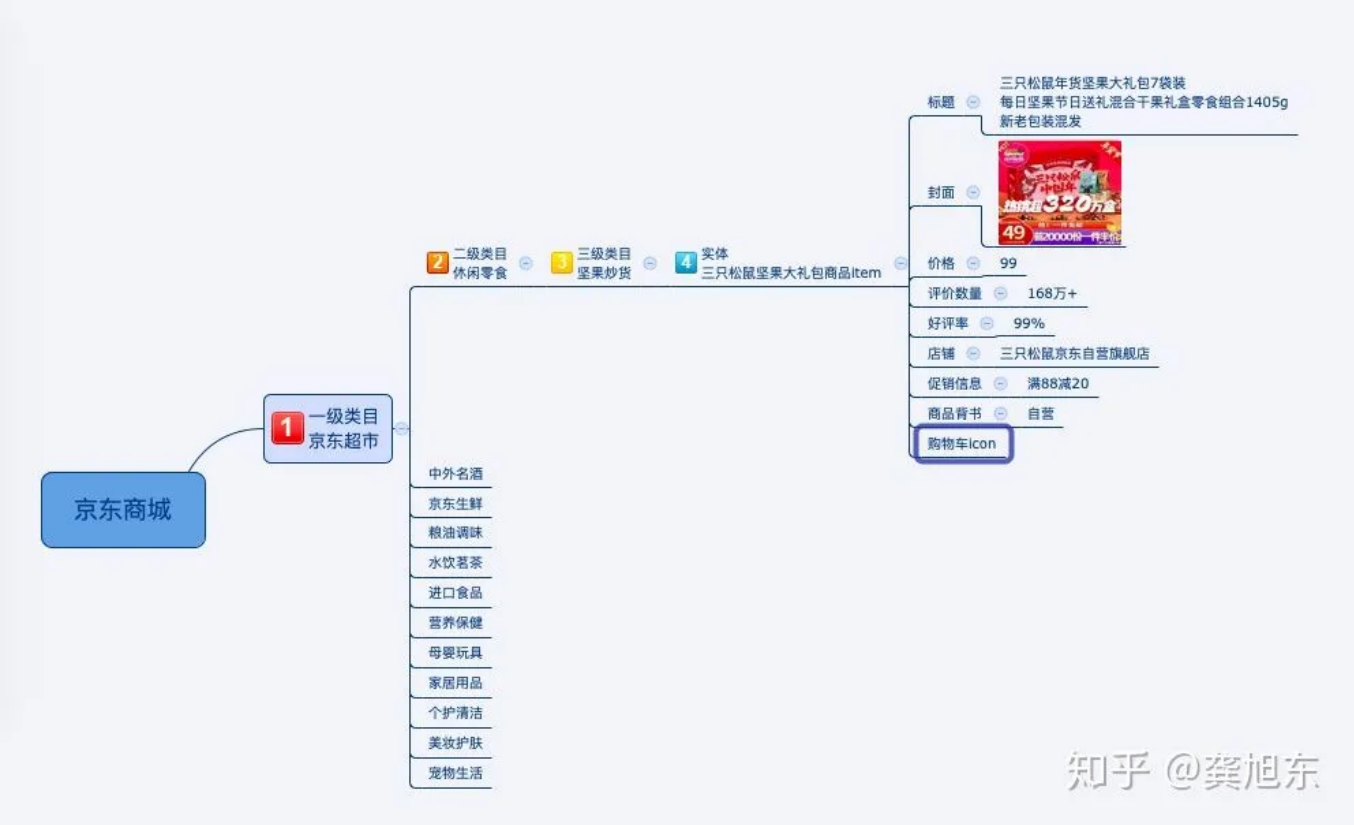
标签体系概览

京东
商品分类

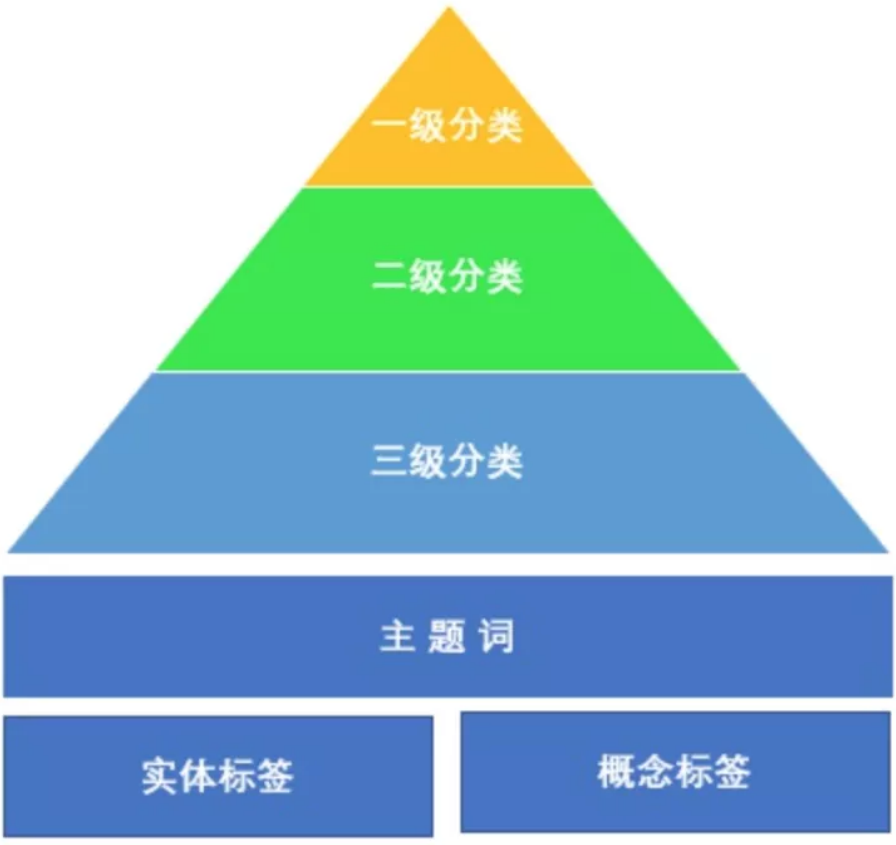
云音乐
歌单分类

头条
资讯分类

以京东的标签体系中的京东超市为例用思维导图来拆解，后面我们会详细的介绍如何构建标签体系。



这里对京东超市标签拆解粒度到三只松鼠年货大礼包的实体级别，实际上各个公司的标签体系大致都是如下构成：



一、二、三级分类体系都很好理解，参考京东超市的拆解，相信大家就会明白。标签体系中实体标签和概念标签不好理解。

实体标签

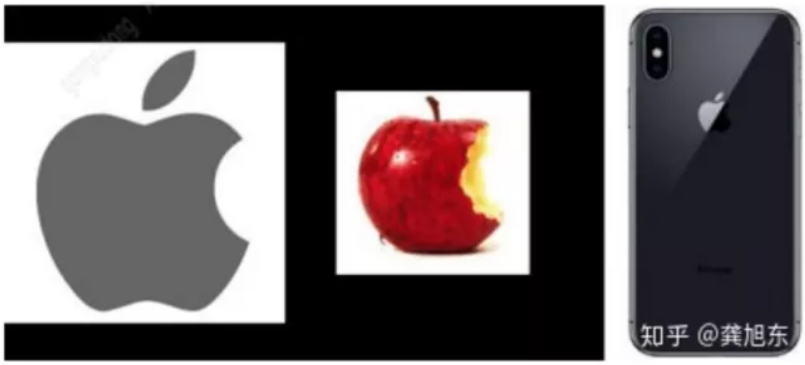
必须是名词，且必须是唯一指代。

学术性的解释逼格高，但是不容易理解，回答下面的问题

- 老板问：苹果，是实体标签吗？
- 给你三秒钟思考
- 你回答：是！
- 老板说：错！
- 你懵逼：靠！为啥不是？

实体标签的要求：名词，且唯一指代。

苹果，是名词，但不是唯一指代，苹果 = 科技公司、手机、水果、牛仔裤



例子	名词	唯一指代	实体标签
苹果	是	不是	不是
苹果手机	是	是	是 知乎 @龚旭东

概念标签

难道我就不能用“苹果”了吗？当然可以用，只不过要给它另外起个名字：概念标签。

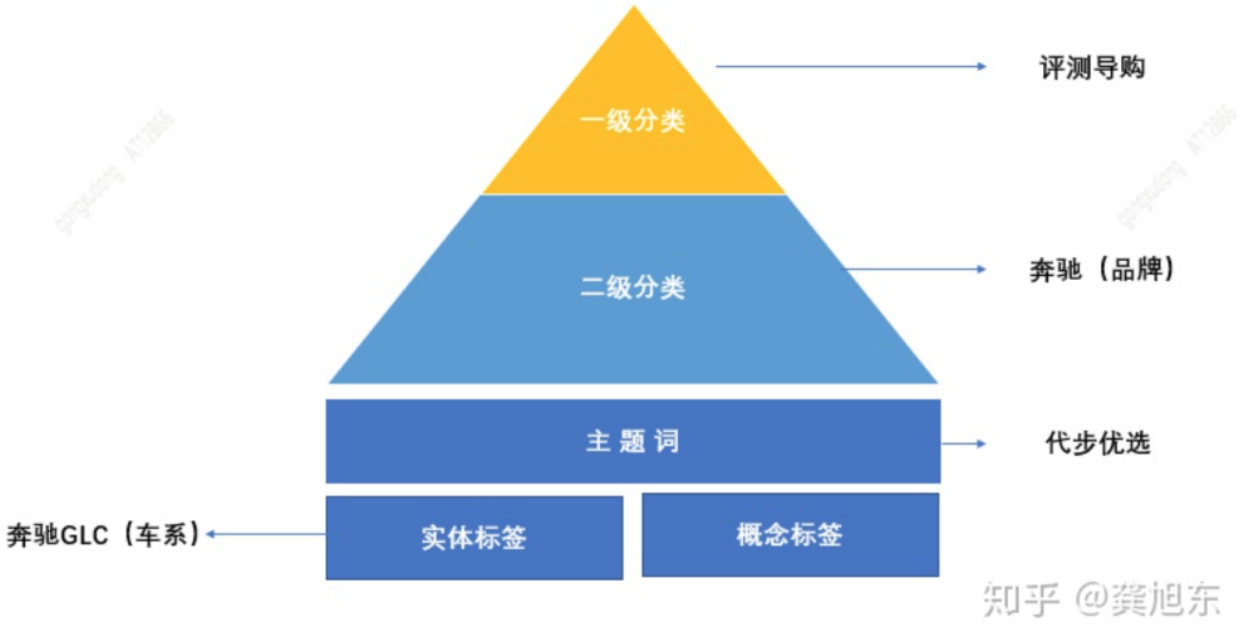
概念标签通常表示的是“一类”或“某种相似”的内容，例如

例子	实体标签	概念标签
苹果	不是	是
苹果手机	是	不是

知乎 @龚旭东

主题词

这里以之家的标签体系举例，要给买车用户推荐评测导购（一级）的文章，用户画像中车的品牌（二级）偏好太粗，而实体标签如奔驰GLC又太细，填补这中间的粒度空白，满足用户购车意图的画像，就加入“代步优选”的主题词，这样不仅保持了推荐的多样性，又不至于过分精准而导致的极度收敛。



以上大致介绍了一下标签体系，那么我们接下介绍一下如何构建标签体系以及其构建过程中应遵循的一些原则。

标签体系构建原则

原则一、放弃大而全的框架，以业务场景倒推标签需求

原则二、标签生成自助化，解决效率和沟通成本

原则三、有效的标签管理机制

分别解释以下为什么提炼出这三个原则，分别用于解决什么问题？

关于第一项原则：

每个公司的产品、运营、商务对标签的诉求有较大的差异，同时不同的运营团队的诉求也存在很大差异，大而全的标签框架实际是站在用户视角搭建的，但是标签的真正应用者是业务方，所以应该从业务视角来实现。

因此最佳的处理方式是，**我们应该放弃顶层的用户抽象视角，针对各业务线或部门的诉求和实际的应用场景，分别将标签聚类起来提供给相应部门。**

之家就是非常典型的情况，商业同学更关心用户的消费能力相关的标签；自驾游负责同学更关心用户的位置和出行相关的标签；车友圈的同学更关注用户的社交活跃相关的标签；所以不可能一套标签覆盖整个运营团队， 这种以业务场景倒推标签需求的方法，能够与业务场景贴合更紧密，可用性上升。

关于第二项原则：

1.标签生成的自助化能够让沟通成本降最低。前面讲到各业务线对标签的定义的理解不同，需要标签系统建设团队花费大量的时间沟通。如果能够让业务方自己定义规则，这必然是沟通成本最低的方式。

2.标签生成的自助化，可重复修改的规则，降低无效标签的堆积。业务一直在发展，如果规则一成不变则很难跟上业务节奏的变化。我曾拜访过一家电商，他们发现半年前定义“母婴客户群”的转化率一直在降低，因此根据实际情况重新修改和定义了“母婴客户群”规则，并命名为“母婴客户群（新）”，这时之前的规则是无效的，且会一直占据计算资源……诸如此类，如果支持规则重复修改的话，这一类无效标签就会大量地消失。

3.释放数据团队人力，释放业务团队的想象力。数据团队应该花较多的精力在企业的整个数据中台或新业务模型方面，而不是处理各业务线的标签诉求和标签维护上，自动化的标签生成能够极大限度地节省人力和释放团队想象力。

关于第三项原则：

1.规则及元信息维护：标签相关的规则和元信息要尽可能的暴露给使用者，让使用者在使用的时候，能清楚知道标签的规则是什么、创建者是谁、维护者是谁、标签的更新频率周期等，而不是没有规则，或者将规则存在标签建设团队内部的一个 word 文档中。

2.调度机制及信息同步：标签之间有一些关联，标签之间的链条断裂，是否有个调度机制或者信息同步机制让大家的工作不被影响。

3.高效统一的输出接口：将所有的业务信息和用户数据信息汇总在一起，有统一的输出接口，改变之前需要针对不同的业务系统开发不同接口的情况。

我们回顾标签体系构建的三原则，本质上是**解决了价值、手段、可持续性**三方面的问题：以业务场景倒推需求，让业务方用起来作为最终目标，让标签系统价值得以实现；标签生成的自助化，它解决的是我们用什么样的手段去实现价值；有效的标签管理机制，意味着一套标签体系能否可持续性地在一家企业里面运作下去。

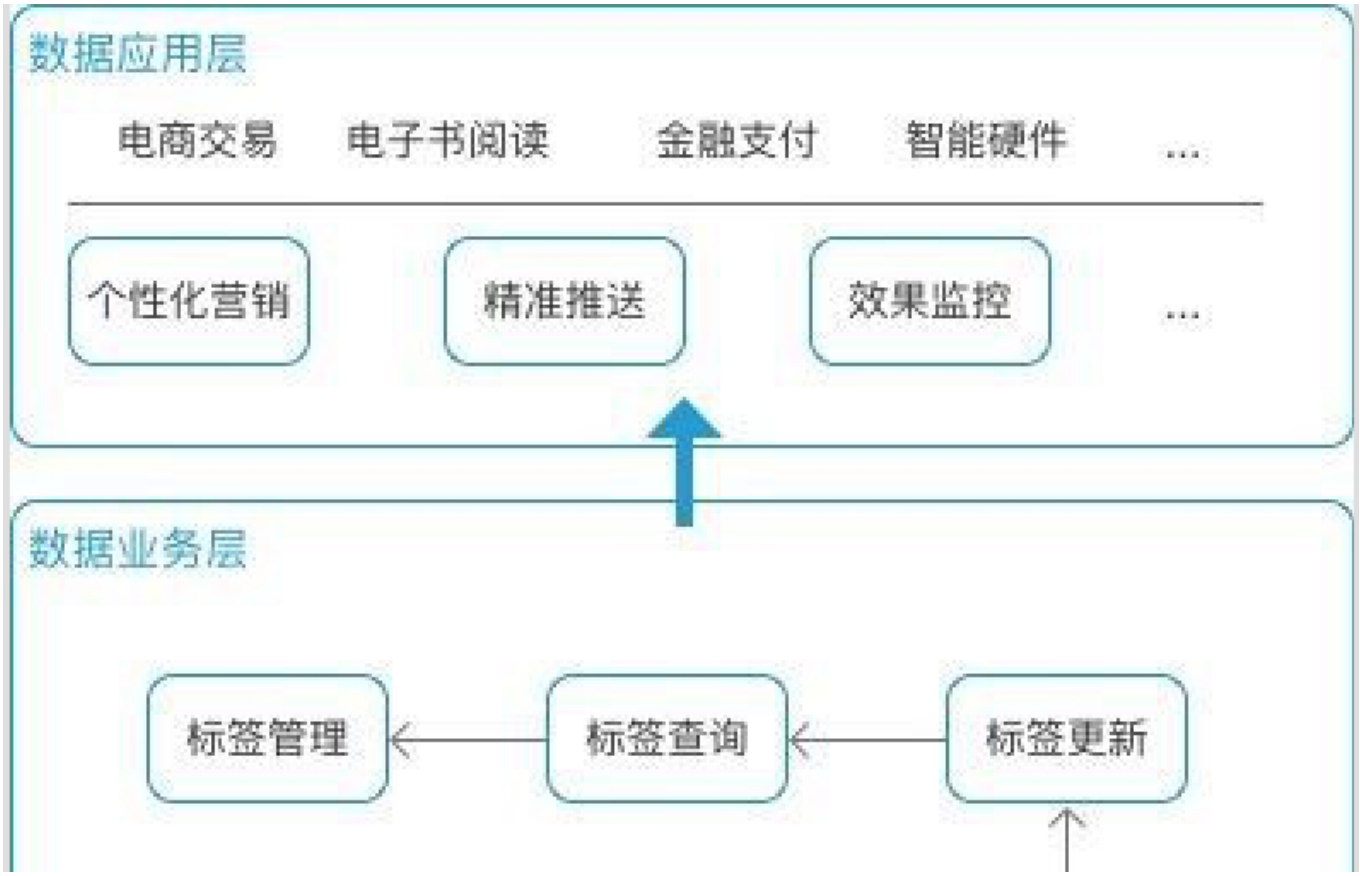
总之，对企业最重要的是：**一套标签系统能不能在业务上用起来，能不能覆盖更广泛的需求，而不是一个大而全的框架。**

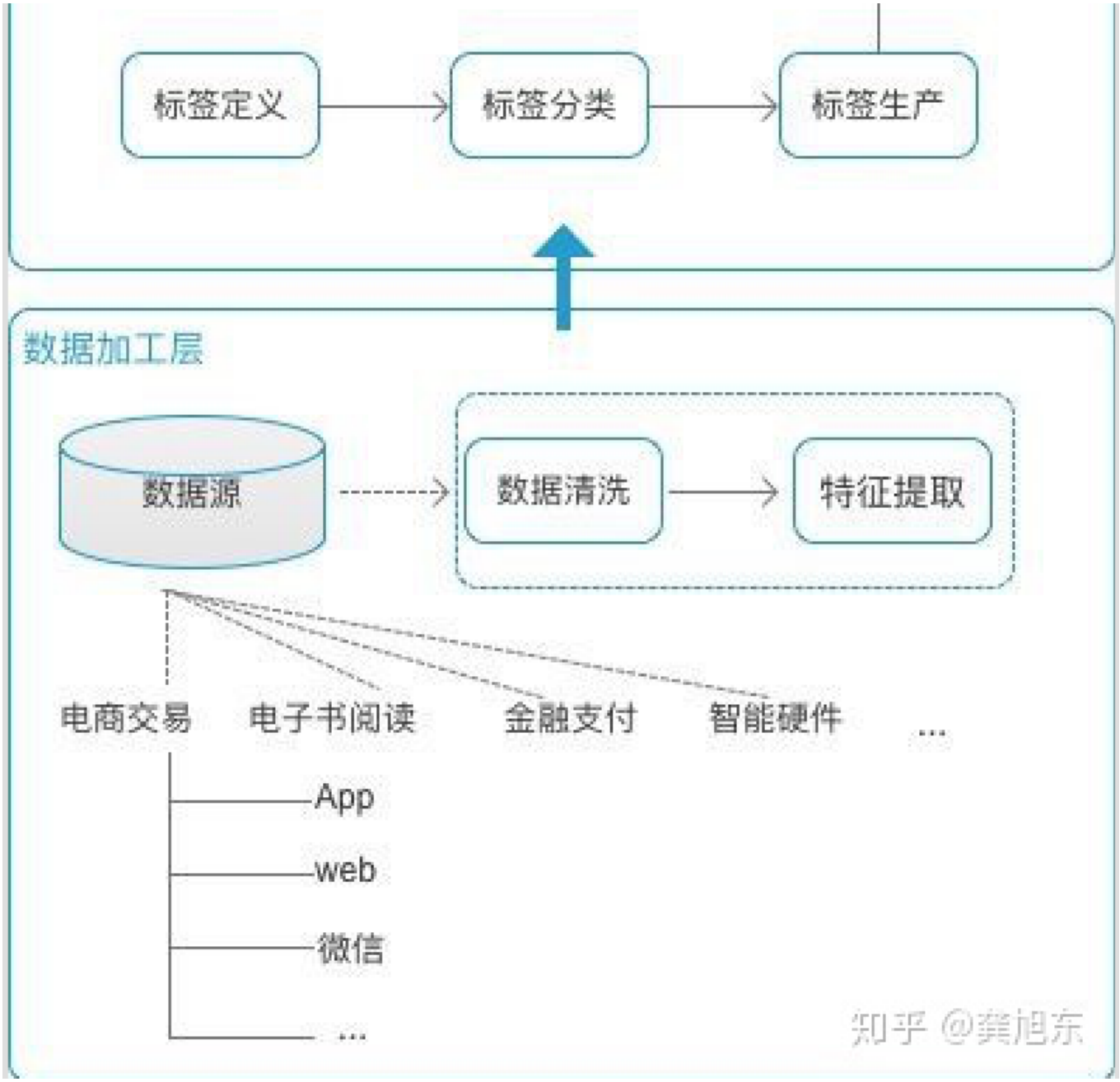
标签体系构建的方法

标签体系的实施架构

标签体系架构可以分为三个部分：数据加工层，数据服务层，数据应用层。每个层面面向用户对象不一样，处理事务有所不同。层级越往下，与业务的耦合度就越小。层级越往上，业务关联性就越强

以某电商公司为例





数据加工层。数据加工层收集，清洗和提取来处理数据。M公司有多个产品线：电商交易，电子书阅读，金融支付，智能硬件等等。每个产品线的业务数据又是分属在不同位置。为了搭建完善的用户标签体系，需要尽可能汇总最大范围内的数据。同时每个产品线的也要集合所有端的数据，比如：App，web，微信，其它第三方合作渠道。

收集了所有数据之后，需要经过清洗：去重，去刷单数据，去无效数据，去异常数据等等。然后再是提取特征数据，这部分就要根据产品和运营人员提的业务数据要求来做就好。

数据业务层。数据加工层为业务层提供最基础数据能力，提供数据原材料。业务层属于公共资源层，并不归属某个产品或业务线。它主要用来维护整个标签体系，集中在一个地方来进行管理。

在这一层，运营人员和产品能够参与进来，提出业务要求：将原材料进行切割。主要完成以下核心任务：

- 定义业务方需要的标签。
- 创建标签实例。
- 执行业务标签实例，提供相应数据。

数据应用层。应用层的任务是赋予产品和运营人员标签的工具能力，聚合业务数据，转化为用户的枪火弹药，提供数据应用服务。

业务方能够根据自己的需求来使用，共享业务标签，但彼此业务又互不影响。实践中可应用到以下几块：

- 智能营销
- Feed流推荐
- 个性化消息push

标签体系的设计

1.业务梳理

以业务需求为导向，可以按下面的思路来梳理标签体系：

- 有哪些产品线？产品线有哪些来源渠道？一一列出。
- 每个产品线有哪些业务对象？比如用户，商品。
- 最后再根据对象聚合业务，每个对象涉及哪些业务？每个业务下哪些业务数据和用户行为？

结果类似如下：

用户	基本特性	年龄； 性别； 生日； 家庭住址； 工作地址； 民族； 消费能力； 支付偏好； 上网设备； 职业； ...
	电商业务	购买金额； 订单数； 收藏商品； 最近浏览； 商品浏览时长； 取消订单数； 退货订单数； ...
	金融支付	支付金额； 支付笔数； 支付方式； 支付时间； ...
	智能硬件	使用时长； 购买时间； 使用次数； ...
	其它业务	...

2.标签分类

按业务需求梳理了业务数据后，可以继续按照业务产出对象的属性来进行分类，主要目的：

- 方便管理标签，便于维护和扩展。
- 结构清晰，展示标签之间的关联关系。
- 为标签建模提供子集。方便独立计算某个标签下的属性偏好或者权重。

梳理标签分类时，尽可能按照MECE原则，相互独立，完全穷尽。每一个子集的组合都能覆盖到父集所有数据。标签深度控制在四级比较合适，方便管理，到了第四级就是具体的标签实例。

一级标签	二级标签	三级标签	四级标签（标签实例）	规则定义	标签类型
人口属性	基本信息	性别	性别-男	系统标注	事实标签
			性别-女	系统标注	事实标签
			性别-未知	系统标注	
		年龄	年龄-xx岁	系统标注	事实标签
		生日	生日-xx	实名认证获取	事实标签
		星座	星座-xx	根据生日-星座得到	事实标签
行为属性	上网习惯	终端类型	终端类型-anroid	系统标注	事实标签
			终端类型-iOS	系统标注	事实标签
		活跃情况	活跃情况-核心用户	满足其中条件之一即视为活跃用户： 1. 过去30天内，发生a行为至少3次。 2. 过去30天内，发生b行为至少3次。 3. 过去30天内，发生c行为至少3次。	模型标签
			活跃情况-活跃用户	满足其中条件之一即视为活跃用户： 1. 过去30天内，发生a行为1-2次。 2. 过去30天内，发生b行为1-2次。 3. 过去30天内，发生c行为1-2次。	模型标签
			活跃情况-新用户	从未进行与业务相关的操作： 1. 行为a 2. 行为b 3. 行为c	模型标签
			活跃情况-老用户	帐号开通以来，发生以下之一的业务： 1. 发生a行为至少1次。 2. 发生b行为至少1次。 3. 发生c行为至少1次。	模型标签
			活跃情况-流失用户	属于老用户，但不符合以下条件之一： 1.过去 30 天时间里，发生a行为 1 次。 2.过去 30 天时间里，发生b行为 1 次。	模型标签
			活跃情况-微信48小时活跃粉丝	符合微信活跃条件，48小时进行以下操作： 1. 新关注 2. 点击自定义菜单 3. 发送消息 4. 扫描二维码 5. 支付成功 6. 用户维权	事实标签
用户分类	人群属性	年龄阶段	年龄阶段-80后	出生时间：1980-1989	事实标签
			年龄阶段-90后	出生时间：1990-1999	事实标签
		地区分布	地区分布-xx	选择城市	事实标签
商业属性		电商业务	购买频度-高频用户	过去12月内，累计订单数超过24	模型标签
			购买频度-中频用户	过去12月内，累计订单数5~24	模型标签
			购买频度-低频用户	过去12月内，累计订单数小于5	模型标签
			购买频度-新用户	至今，累计订单数为0	模型标签
		金融支付	支付频度-高频用户	过去30日内，累计支付笔数大于150	模型标签
			支付频度-中频用户	过去30日内，累计支付笔数在20-150	模型标签
			支付频度-低频用户	过去30日内，累计支付笔数小于20	模型标签
			支付频度-新用户	至今，支付笔数为0	模型标签
			消费订单比例-消费狂	消费订单比例高于60%或过去30日内，超过30件	模型标签
			消费订单比例-消费达人	消费订单比例达到在20-60%或过去30日内，在10-30件之间	模型标签
			消费订单比例-普通者	电商订单比例达到低于10%或过去30日内低于10件	模型标签
		充值	充值-充值新用户	至今，未充过值	模型标签
			充值-土豪	过去12个月，累计充值超过1500	模型标签
			充值-充值大户	过去12个月，累计充值在200-1500之间	模型标签
			充值群众	过去12个月，累计充值低于1500	模型标签
		优惠券	优惠券-敏感度高用户	过去6个月，优惠券使用率超过50%	模型标签
			优惠券-敏感度中用户	过去6个月，优惠券使用率在10-50%	模型标签
			优惠券-敏感度低用户	过去6个月，优惠券使用率低于10%	模型标签
		积分值	积分-等级高用户	积分值超过xxx	模型标签
			积分-等级中用户	积分值在xxx-xxx之间	模型标签
			积分-等级低用户	积分值低于xxx	模型标签
		品牌偏好	品牌偏好-高端	过去12个月，买过m类产品占比超过50%	模型标签
			品牌偏好-中端	过去12个月，买过b类产品占比超过50%	模型标签
			品牌偏好-低端	过去12个月，买过c类产品占比超过50%	模型标签

消费习惯	支付偏好	支付偏好-微信	最近3个月里，微信支付占比
		支付偏好-支付宝	最近3个月里，支付宝支付占
		支付偏好-钱包	最近3个月里，钱包支付占比

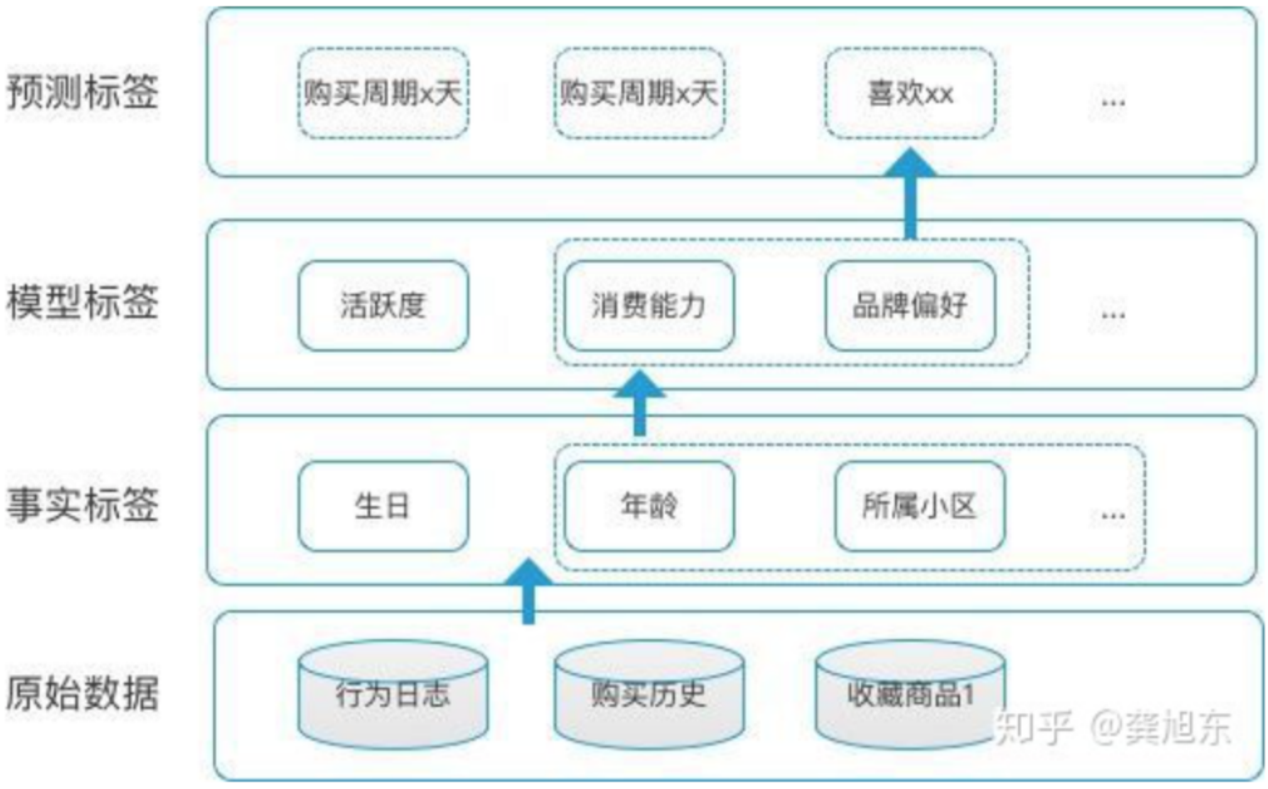
3.标签的模型

按数据的实效性来看，标签可分为

- 静态属性标签。长期甚至永远都不会发生改变。比如性别，出生日期，这些数据都是既定的事实，几乎不会改变。
- 动态属性标签。存在有效期，需要定期地更新，保证标签的有效性。比如用户的购买力，用户的活跃情况。

从数据提取维度来看，标签数据又可以分为类型。

- 事实标签。既定事实，从原始数据中提取。比如通过用户设置获取性别，通过实名认证获取生日，星座等信息。
- 模型标签。没有对应数据，需要定义规则，建立模型来计算得出标签实例。比如支付偏好度。
- 预测标签。参考已有事实数据，来预测用户的行为或偏好。比如用户a的历史购物行为与群体A相似，使用协同过滤算法，预测用户a也会喜欢某件物品。



4.标签的处理

为什么要从两个维度来对标签区分？这是为了方便用户标签的进一步处理。

静态动态的划分是面向业务维度，便于运营人员理解业务。这一点能帮助他们：

- 理解标签体系的设计。
- 表达自己的需求。

事实标签，模型标签，预测标签是面向数据处理维度，便于技术人员理解标签模块功能分类，帮助他们：

- 设计合理数据处理单元，相互独立，协同处理。
- 标签的及时更新及数据响应的效率。

以上面的标签图表为例，面临以下问题：

- 属性信息缺失怎么办？比如，现实中总有用户未设置用户性别，那怎么才能知道用户的性别呢？
- 行为属性，消费属性的标签能不能灵活设置？比如，活跃运营中需要做A/B test，不能将品牌偏好规则写死，怎么办？
- 既有的属性创建不了我想要的标签？比如，用户消费能力需要综合结合多项业务的数据才合理，如何解决？

模型标签的定义解决的就是从无到有的问题。建立模型，计算用户相应属性匹配度。现实中，事实标签也存在数据缺失情况。

> 比如用户性别未知，但是可以根据用户浏览商品，购买商品的历史行为来计算性别偏好度。当用户购买的女性化妆品和内衣较多，偏好值趋近于性别女，即可以推断用户性别为女。

模型计算规则的开放解决的是标签灵活配置的问题。运营人员能够根据自己的需求，灵活更改标签实例的定义规则。比如图表中支付频度实例的规则定义，可以做到：

- 时间的开放。支持时间任意选择：昨天，前天，近x天，自定义某段时间等等。
- 支付笔数的开放。大于，等于，小于某个值，或者在某两个值区间。

标签的组合解决就是标签扩展的问题。除了原有属性的规则定义，还可以使用对多个标签进行组合，创建新的复合型标签。比如定义用户的消费能力等级。

标签最终呈现的形态要满足两个需求：

- 标签的最小颗粒度要触达到具体业务事实数据，同时支持对应标签实例的规则自定义。
- 不同的标签可以相互自由组合为新的标签，同时支持标签间的关系，权重自定义。