一、前言

数据仓库具有面向主题的特性，那么就会有主题的概念，数仓建设是遵循纵向分层开发，横向划分主题域设计，数仓分层就不在这次谈了，**这次我会结合本人数仓工作实践总结的经验来聊聊数仓主题域划分，同时会引申出主题划分，和数据域是什么，业务过程等。**

这个对于大数据数仓工程师来说是必备的能力，比如当你面临着一个新业务的开启，需要从0到1开始搭建数据仓库或者数据集市，这时候就要考虑到主题域和主题的合理划分。

当然本次分享的内容都是从个人实际出发，有疑问或者反馈可以通过关注公众号留言共同探讨，感谢关注。

二、数仓建设的步骤

- **1. 业务调研**
  数仓开发侧是承上对接业务研发侧&承下对接数据分析侧，在数仓建设前期要对上游业务过程和对下游数据分析指标体系有所了解和熟知，然后拉齐上下游沟通数据口径和数仓搭建。
- **2. 主题域划分**
- **3. 主题划分**
- **4. 输出总线矩阵**
  即业务过程和维度，组建成的矩阵
- **5. 数仓分层设计模型表**
- **6. 数仓公共层表迭代升级**

三、主题和主题域

下面结合本人对搬家业务的数仓建设，进行主题域划分和主题划分实践，当然项目的大小决定着这是一个小型的数据集市还是 企业级的数据仓库。

主题域的划分

数仓主题域：主题域通常是联系较为紧密的数据主题的集合，根据业务需求分析的视角进行划分抽象归类。

划分方法：主题域划分的方法一般有几种

- 要么按照业务过程来划分，一个业务过程抽象出一个主题域，比如业务系统中的商品、交易、物流 等
- 要么按照业务部门来划分，一个业务部门抽象出一个主题域，比如中台部门、业务运营部门、供应链部门 等
- 要么按照业务系统来划分，一个业务系统抽象出一个主题域，比如搬家系统、erp系统 等

主题的划分

数仓主题：是在较高层次上将企业生产上的各个系统中某一分析对象 的数据进行整合、归类并分析的一种范围，属于一个抽象概念，简单点说每一个主题对应一个宏观分析领域。

划分方法：说白了主要就是要识别出分析对象主体，做主题划分和主题域划分，个人建议是要站在全局的视角来看，然后先划分出主题域，再接着在主题域里面划分出各个主题，主题域的划分一般比较谨慎，一旦定下来了避免频繁变动，虽然数仓建设是迭代建设的，不能保证一次性初始化好，但我们的主题域划分和主题划分要尽可能地涵盖企业的所有业务，以及在新业务进来时能够无影响地被包含进来和可扩展主题域。

个人案例实践分享

我就分享我负责过的搬家业务数仓建设中，我是如何划分主题域和划分主题的，规模相当于数据集市，即小型的数据仓库

划分主题域：
首先我是按照业务系统来划分的，搬家是企业业务中一个独立的业务线，所对应的业务系统也是跟其他系统是独立开的，那么这时候我按照业务系统来划分，就不会在建设过程中出现一些‘扯皮’操作，出现数据边界归属问题。

划分主题：
上面的主题域划分完了后就产生一个搬家主题域，然后把搬家分析作为一个分析领域，那么‘搬家分析’所涉及到的主要分析对象就有用户、订单、搬运工 等，则数仓的主题就可以划分为用户主题、订单主题、搬运工主题 等。

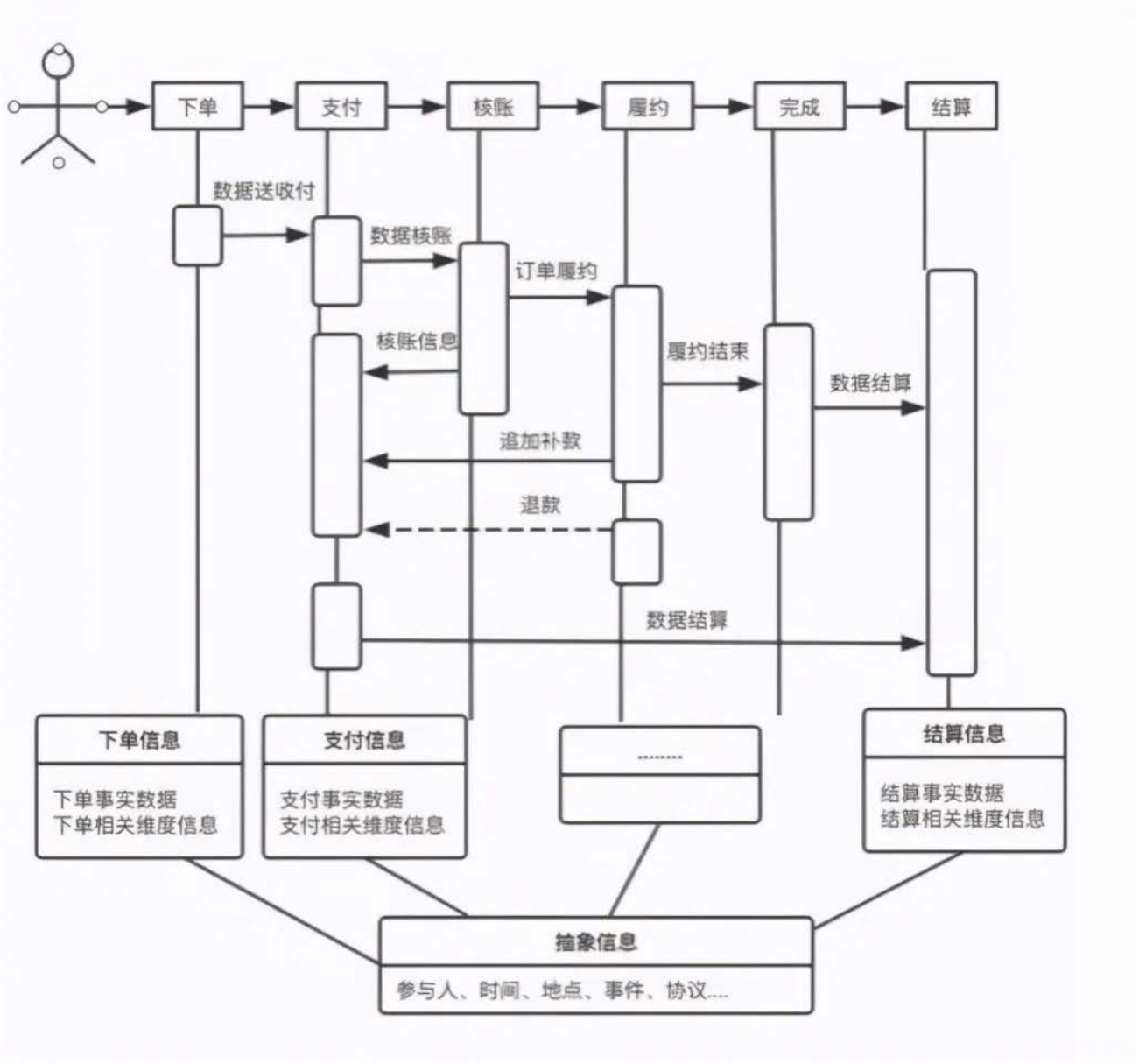
健壮性评估：
当后续搬家主题域业务新增，我还可以轻松地扩展出其他主题，毕竟按照上面的划分法，搬家的数据基本都划分在搬家主题域，剩下的就是搬家有新业务进来时扩展新主题或包含进已有主题。

分享业界其他的案例

分享网上搜索到的 马蜂窝数仓主题、主题域划分案例

以马蜂窝订单交易模型的建设为例，基于业务生产总线的设计是常见的模式，首先调研订单交易的完整过程，定位过程中

的关键节点，确认各节点上发生的核心事实信息。



数据域	主题	主题关键字及简写	覆盖内容
交易	订单	Order(ord)	酒店、大交通、自由行、等各类订单
	财务	Finance(fin)	账务、支付、结算
	营销	Marketing(mkt)	市场营销、促销、优惠券
	产品	Product(prd)	平台售卖的产品、商品、SKU等相关属性及供应链如库存等
	客服	Customer Service Center （csc)	呼叫中心、IM等
流量	流量	Flow(flw)	各类终端流量、业务流量、渠道流量等
内容	内容	Community(cmt)	内容5大业务线、目的地、POI等
参与人	用户	User（usr)	平台个人用户
	设备	Device(dvc)	各类设备：手机、ipad等
	商家	Vendor(vdr)	供应商、平台卖家等
	员工	employee（epl)	马蜂窝员工

四、主题域、数据域、业务过程

总是听到数据域，那么数据域和主题域是有什么关系呢，参考《阿里巴巴大数据之路》书籍和网上有人总结过这么一段，如下：

主题域：面向业务过程，将业务活动事件进行抽象的集合，如下单、支付、退款都是业务过程，针对公共明细层（DWD）进行主题划分。

数据域：面向业务分析，将业务过程或者维度进行抽象的集合，针对公共汇总层（DWS）进行数据域划分。

业务过程：指企业的业务活动事件，如下单、支付、退款都是业务过程，业务过程就是一个不可拆分的行为事件。

其实数据域跟主题域的差别不大，很大情况下两者就等同于一个概念的。

表 9.1　名词术语解释

名词术语	解　　释
数据域	指面向业务分析，将业务过程或者维度进行抽象的集合。其中，业务过程可以概括为一个个不可拆分的行为事件，在业务过程之下，可以定义指标；维度是指度量的环境，如买家下单事件，买家是维度。为保障整个体系的生命力，数据域是需要抽象提炼，并且长期维护和更新的，但不轻易变动。在划分数据域时，既能涵盖当前所有的业务需求，又能在新业务进入时无影响地被包含进已有的数据域中和扩展新的数据域
业务过程	指企业的业务活动事件，如下单、支付、退款都是业务过程。请注意，业务过程是一个不可拆分的行为事件，通俗地讲，业务过程就是企业活动中的事件
时间周期	用来明确数据统计的时间范围或者时间点，如最近 30 天、自然周、截至当日等
修饰类型	是对修饰词的一种抽象划分。修饰类型从属于某个业务域，如日志域的访问终端类型涵盖无线端、PC 端等修饰词
修饰词	指除了统计维度以外指标的业务场景限定抽象。修饰词隶属于一种修饰类型，如在日志域的访问终端类型下，有修饰词 PC 端、无线端等
度量 / 原子指标	原子指标和度量含义相同，基于某一业务事件行为下的度量，是业务定义中不可再拆分的指标，具有明确业务含义的名词，如支付金额
维度	维度是度量的环境，用来反映业务的一类属性，这类属性的集合构成一个维度，也可以称为实体对象。维度属于一个数据域，如地理维度（其中包括国家、地区、省以及城市等级别的内容）、时间维度（其中包括年、季、月、周、日等级别的内容）
维度属性	维度属性隶属于一个维度，如地理维度里面的国家名称、国家 ID、省份名称等都属于维度属性