基本概念

数据仓库概念由世界公认的数据仓库之父Bill Inmon（比尔·恩门）在1991年出版的“Building the Data Warehouse”（《建立数据仓库》）中提出：

数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。

数据建模是将数据进行有序、有结构地分类组织和存储的方法。其思路是**从企业分析决策等需求出发**，建设一个数据易用、产出稳定、质量可信、指标丰富、能够提供标准化数据服务的大数据体系。该体系同时具有高扩展、强复用、低成本的特性，能够避免数据重复建设和指标冗余建设，保障数据的规范性、指标的一致性，能够有效地管理和控制日益增长的存储和计算消耗。

目前工业界流行的建模方法主要是维度建模，维度模型是由数仓领域大师Ralph Kimball所倡导的，大师著作的《The Data Warehouse Toolkit-The Complete Guide to Dimensional Modeling》是数仓工程领域最流行的维度建模经典。

数仓建设体系

有句话说得好，一个好的架构已经成功了80%。同样一个好的建模体系能释放计算消耗、存储消耗、人力资源、时间成本等。同时好的架构还有助于提供标准化的、高效的、优质的、稳定的数据服务。所以，在我们具体建模之前，不要着急动手，要先规划好我们的数据体系架构，划定边界，搭建统一规范的体系架构。

1）确定建模技术平台

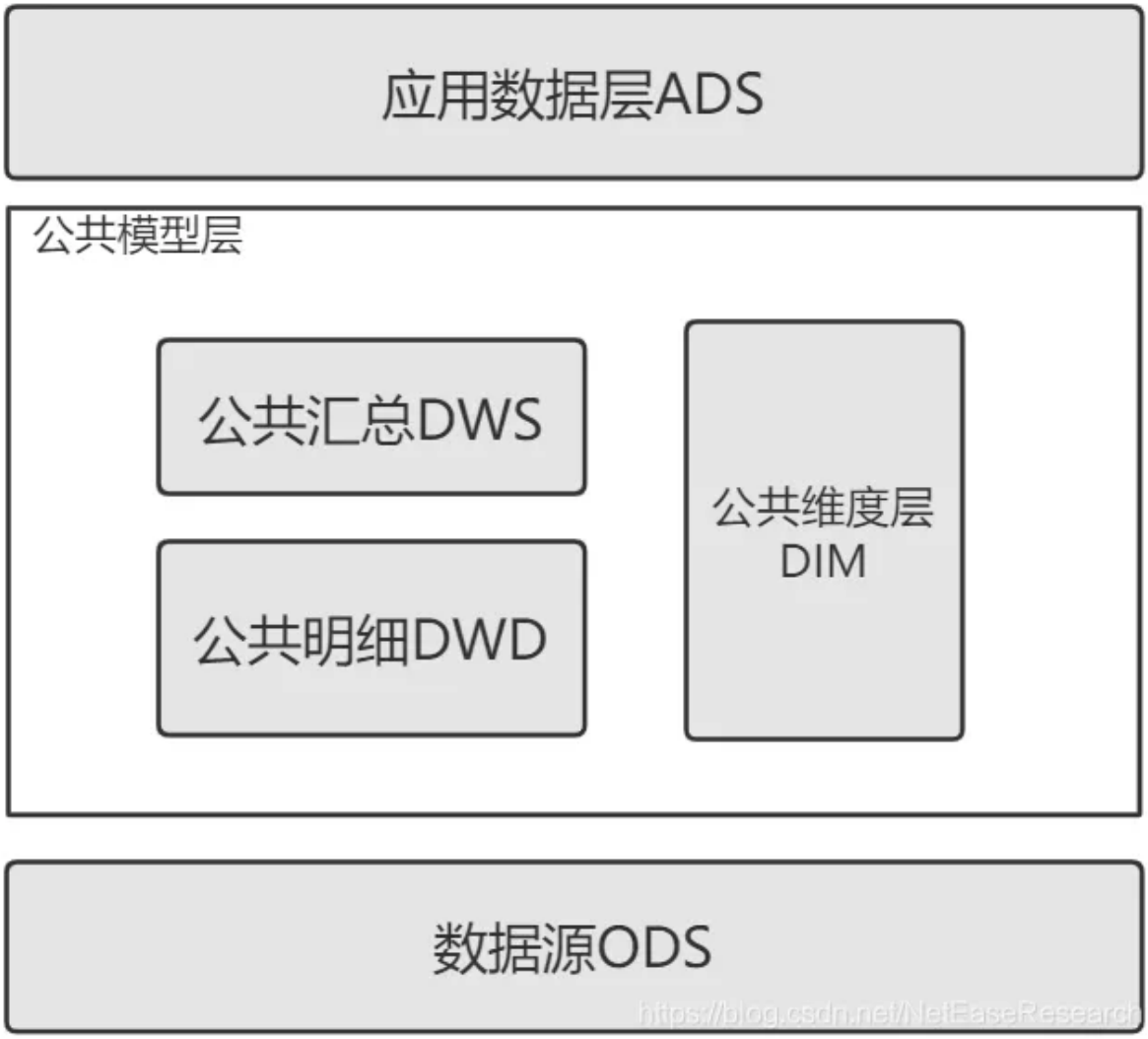
建设数仓首先是需要一个好用的数据平台，网易内部用的是网易数帆旗下的网易易数，一个提供数据开发、数据治理、数据分析及可视化、数据服务、数据应用的一站式全链路数据生产力平台，有易研发、易治理、易服务、易应用的特点。

所谓好用的工具事半功倍。网易提倡“人人用数据、天天用数据”，依托于体系化的数据中台产品，搭建标准化的数仓体系。

2）定好模型层次

数据从产出原生数据到最后生成指标的过程，就像建房子要明确地基、一楼、二楼的作用一样，数据模型同样需要定义好模型层次，规划好每个层次该做什么事情，包括：明细层、轻度汇总层、主题层、应用层等。

目前应用广泛的是如下三层数据模型：



（1）操作数据层(ODS)**：**全称Operation Data Store，我们不是数据的生成者，我们是数据的搬运工，ODS层存储的就是最原始的数据，是我们即将搬运清洗的数据。

（2）公共模型层，包含：

- **DWD层(data warehouse detail)**：**这层存放的是明细数据，**面向业务过程建模。一方面对ODS层的数据清洗、规范化，保证数据的一致性。比如空值处理、时间格式统一、异常值处理、枚举值统一、脱敏、编码转换等。另一方面，DWD层允许退化一些常用维度在里面，减小和维表的关联次数，提升明细数据表的易用性。
- **DWS层(data warehouse service)：**公共汇总数据层，**面向分析主题建模**，为上层数据产品、分析、应用和服务等提供公共计算指标。
- **DIM（dimension）**：**公共维度表，用于建立**一致性维度数据**，规范化维度属性，降低数据计算口径和算法不一致风险。

（3）应用数据层ADS**：**全称Application Data Store，**面向应用汇总数据**，存放的是基于应用组装的数据或者产品个性化标签数据，面向实际需求，如报表等。

划分模型层次的目的是定一个标准，就像大型商场去4楼吃饭去5楼看电影，想看明细数据就去DWD层，想看维度就去DIM层。

3）梳理业务过程和维度

先来理解下业务过程、度量、维度、维度属性的含义。

- **业务过程：**指企业活动中的具体的、不可拆分的行为事件，如播放、下载、分享等都是业务过程。**业务过程一般体现在****DWD层模型中****。**
- **度量：**对某一业务过程行为的度量，也称原子指标，不能继续拆分。如播放业务过程中，播放次数、播放人数等
- **维度：**维度是实体对象，描述的是度量的环境，是我们观察业务的角度。反应业务的一类属性，这类属性集合构成一个维度，比如，播放过程，度量是播放次数，维度是用户、歌曲、设备、地理、时间等。一个维度落地就是一张维表。
- **维度属性：**一个维度里面的具体属性，如用户维度的用户昵称、年龄、性别等。通俗讲就是维表的列。

搭建一个稳定的数据体系，我们需要了解数据，知道库里有什么，就需要对我们的产品功能和业务进行全盘摸底，用建模的术语就是要梳理清楚全产品所有或者重点核心业务过程及其涉及到的所有可能维度，对部门数据进行全盘了解。输出的形式是业务过程×维度矩阵，如：

务过程	**一致性维度**				
歌曲	用户	设备	时间	地理	
播放歌曲	√	√	√	√	√
流量点击	×	√	√	√	√

4）划分数据域

（1）数据域概念

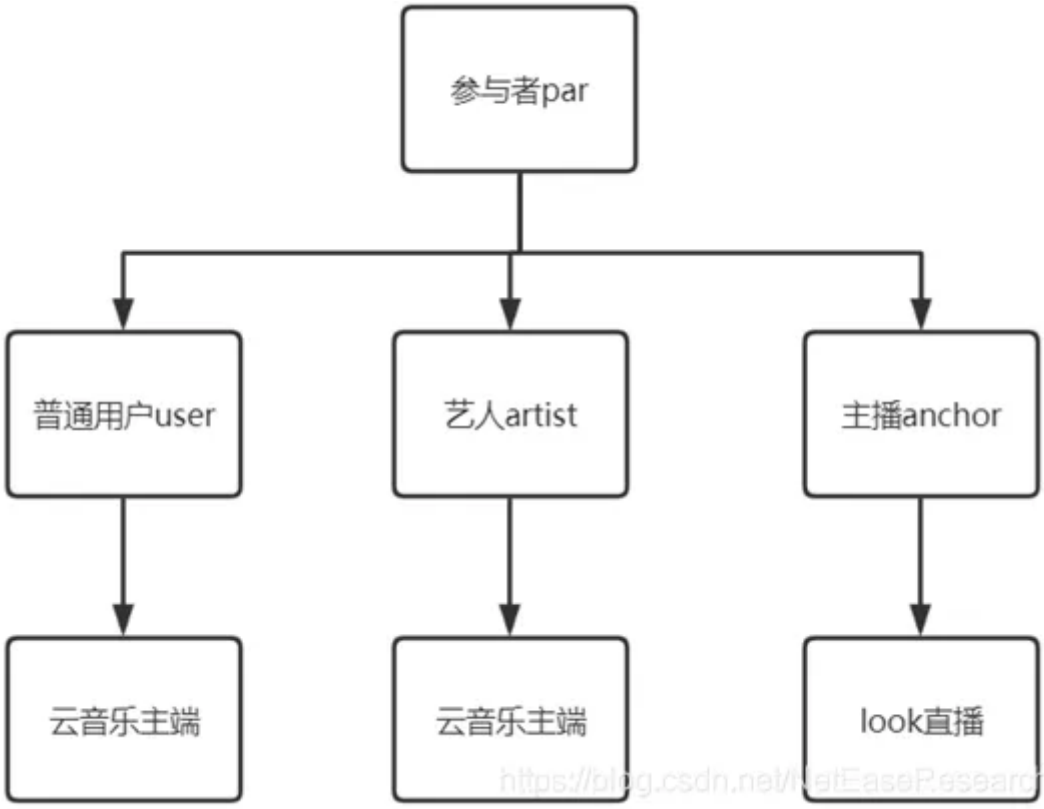
数据仓库是**面向主题**的应用。主题是一个**抽象概念**，是在较高层次上将企业数据进行综合、归类和分析利用的一个抽象，比如用户主题、营收主题、社交主题等。主题细分的话可以分成业务主题和分析主题，**业务主题面向业务过程，分析主题面向汇总层。**

建模除了需要横向的分层外，通常也需要根据业务特点纵向的划分数据域。**数据域**是指面向**业务分析**的，是联系较为紧密的数据主题的集合，是对业务对象高度概括的抽象归类，数据域有时候也可以称为主题域。

（2）数据域层级

数据域也是有层级的，简单业务只要一层数据域，复杂业务需要两层，如果业务板块比较多的话，可能需要三层。

以网易云音乐用户为例，用户身份有：普通用户、音乐人、艺人、创作者、主播等，用户包含业务板块有：主端云音乐、直播、k歌等。那数据域可以划分三级：参与者-用户身份-具体业务板块。



（3）数据域划分方法

数据域的划分可以先梳理业务过程、维度或者产品功能模块，了解现有的需求，如分析报表等，然后抽象到业务主题、分析主题，最后再整合、归类、划分、反复推敲、确认。

（4）数据域注意事项

① 数据域不能轻易变动，所以我们在划分数据域的过程中，需要反复推敲和确认。

② 数据域是需要长期维护的，既要确保能涵盖当前所有业务范围，新业务出现时，要么能融合进原有的数据域，要么可以有边界地新增数据域。

③ 数据域边界不能模糊不清，一定要有明确的边界。

④ 一个维度或业务过程只能隶属于某一个数据域，某个数据域可以包含多个维度或业务过程。

按照数据域划分方法划分好数据域后，每个业务过程都有唯一的数据域。那么，业务过程×维度矩阵可以进一步扩展，构建面向业务过程的数据域×业务过程×维度的总线矩阵。

据域	**业务过程**	**一致性维度**				
歌曲	用户	设备	时间	地理		
互动域	播放歌曲	√	√	√	√	√
日志域	流量点击	×	√	√	√	√

5）定好数据规范和开发规范

数仓是团队建设的工作，所谓一千个读者有一千个哈姆雷特，那么最好建立统一标准的数据规范和开发规范，保持指标、表、流、存储等的一致性。

这里的规范涉及到数仓建模所有过程的规范，包括不限于：

- 模型层次调用规范
- 数据域命名规范

- 建表规范
- 临时表、正式表命名规范
- 原子指标、派生指标命名规范
- 数据格式规范
- 数据存储规范
- 作业流规范
- 枚举值规范
- 维度规范
- 词根规范
- 公共字段规范
- 计算指标来源规范
- 指标一致性建设规范
- 交付标准规范