**数据仓库建模备课资料**

目录

[1 什么是数据模型 1](#_Toc435975821)

[2 为什么需要数据仓库模型 2](#_Toc435975822)

[3 如何建设数据仓库模型 3](#_Toc435975823)

[3.1  数据仓库数据模型架构 3](#_Toc435975824)

[3.2 数据仓库建模阶段划分 4](#_Toc435975825)

[3.4 数据仓库建模方法 6](#_Toc435975826)

[4 数据仓库建模案例 9](#_Toc435975827)

[4.1 零售行业案例分析 9](#_Toc435975828)

# 什么是数据模型

数据模型是抽象描述现实世界的一种工具和方法，是通过抽象的实体及实体之间联系的形式，来表示现实世界中事务的相互关系的一种映射。在这里，数据模型表现的抽象的是实体和实体之间的关系，通过对实体和实体之间关系的定义和描述，来表达实际的业务中具体的业务关系。

数据仓库模型是数据模型中针对特定的数据仓库应用系统的一种特定的数据模型，一般的来说，我们数据仓库模型分为几下几个层次，如图 2 所示。

**图 2. 数据仓库模型**



通过上面的图形，我们能够很容易的看出在整个数据仓库得建模过程中，我们需要经历一般四个过程：

* **业务建模**，生成业务模型，主要解决业务层面的分解和程序化。
* **领域建模**，生成领域模型，主要是对业务模型进行抽象处理，生成领域概念模型。
* **逻辑建模**，生成逻辑模型，主要是将领域模型的概念实体以及实体之间的关系进行数据库层次的逻辑化。
* **物理建模**，生成物理模型，主要解决，逻辑模型针对不同关系型数据库的物理化以及性能等一些具体的技术问题。

因此，**在整个数据仓库的模型的设计和架构中，既涉及到业务知识，也涉及到了具体的技术，我们既需要了解丰富的行业经验，同时，也需要一定的信息技术来帮助我们实现我们的数据模型，最重要的是，我们还需要一个非常适用的方法论，来指导我们自己针对我们的业务进行抽象，处理，生成各个阶段的模型。**

# 为什么需要数据仓库模型

在数据仓库的建设中，我们一再强调需要数据模型，那么数据模型究竟为什么这么重要呢？首先我们需要了解整个数据仓库的建设的发展史。

**数据仓库的发展大致经历了这样的三个过程：**

* **简单报表阶段：**这个阶段，系统的主要目标是解决一些日常的工作中业务人员需要的报表，以及生成一些简单的能够帮助领导进行决策所需要的汇总数据。这个阶段的大部分表现形式为数据库和前端报表工具。
* **数据集市阶段：**这个阶段，主要是根据某个业务部门的需要，进行一定的数据的采集，整理，按照业务人员的需要，进行多维报表的展现，能够提供对特定业务指导的数据，并且能够提供特定的领导决策数据。
* **数据仓库阶段：**这个阶段，主要是按照一定的数据模型，对整个企业的数据进行采集，整理，并且能够按照各个业务部门的需要，提供跨部门的，完全一致的业务报表数据，能够通过数据仓库生成对对业务具有指导性的数据，同时，为领导决策提供全面的数据支持。

通过数据仓库建设的发展阶段，我们能够看出，数据仓库的建设和数据集市的建设的重要区别就在于数据模型的支持。因此，数据模型的建设，对于我们数据仓库的建设，有着决定性的意义。

**一般来说，数据模型的建设主要能够帮助我们解决以下的一些问题：**

* **进行全面的业务梳理，改进业务流程。**在业务模型建设的阶段，能够帮助我们的企业或者是管理机关对本单位的业务进行全面的梳理。通过业务模型的建设，我们应该能够全面了解该单位的业务架构图和整个业务的运行情况，能够将业务按照特定的规律进行分门别类和程序化，同时，帮助我们进一步的改进业务的流程，提高业务效率，指导我们的业务部门的生产。
* **建立全方位的数据视角，消灭信息孤岛和数据差异。**通过数据仓库的模型建设，能够为企业提供一个整体的数据视角，不再是各个部门只是关注自己的数据，而且通过模型的建设，勾勒出了部门之间内在的联系，帮助消灭各个部门之间的信息孤岛的问题，更为重要的是，通过数据模型的建设，能够保证整个企业的数据的一致性，各个部门之间数据的差异将会得到有效解决。
* **解决业务的变动和数据仓库的灵活性。**通过数据模型的建设，能够很好的分离出底层技术的实现和上层业务的展现。当上层业务发生变化时，通过数据模型，底层的技术实现可以非常轻松的完成业务的变动，从而达到整个数据仓库系统的灵活性。
* **帮助数据仓库系统本身的建设。**通过数据仓库的模型建设，开发人员和业务人员能够很容易的达成系统建设范围的界定，以及长期目标的规划，从而能够使整个项目组明确当前的任务，加快整个系统建设的速度。

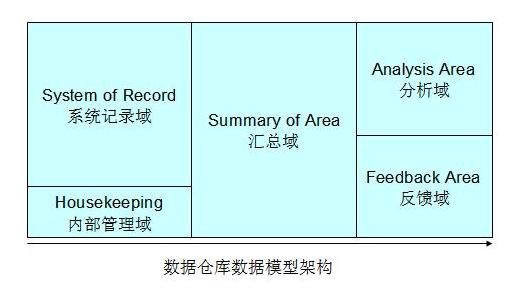
# 如何建设数据仓库模型

建设数据模型既然是整个数据仓库建设中一个非常重要的关键部分，那么，怎么建设我们的数据仓库模型就是我们需要解决的一个问题。这里我们将要详细介绍如何创建适合自己的数据模型。

### 3.1  数据仓库数据模型架构

数据仓库的数据模型的架构和数据仓库的整体架构是紧密关联在一起的，我们首先来了解一下整个数据仓库的数据模型应该包含的几个部分。从下图我们可以很清楚地看到，整个数据模型的架构分成 5 大部分，每个部分其实都有其独特的功能。

**图 3. 数据仓库数据模型架构**



从上图我们可以看出，**整个数据仓库的数据模型可以分为大概 5 大部分**：

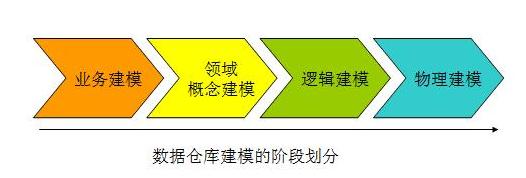
* **系统记录域（System of Record）：**这部分是主要的数据仓库业务数据存储区，数据模型在这里保证了数据的一致性。
* **内部管理域（Housekeeping）：**这部分主要存储数据仓库用于内部管理的元数据，数据模型在这里能够帮助进行统一的元数据的管理。
* **汇总域（Summary of Area）：**这部分数据来自于系统记录域的汇总，数据模型在这里保证了分析域的主题分析的性能，满足了部分的报表查询。
* **分析域（Analysis Area）：**这部分数据模型主要用于各个业务部分的具体的主题业务分析。这部分数据模型可以单独存储在相应的数据集市中。
* **反馈域（Feedback Area）：**可选项，这部分数据模型主要用于相应前端的反馈数据，数据仓库可以视业务的需要设置这一区域。

通过对整个数据仓库模型的数据区域的划分，我们可以了解到，一个好的数据模型，不仅仅是对业务进行抽象划分，而且对实现技术也进行具体的指导，它应该涵盖了从业务到实现技术的各个部分。

### 3.2 数据仓库建模阶段划分

我们前面介绍了数据仓库模型的几个层次，下面我们讲一下，针对这几个层次的不同阶段的数据建模的工作的主要内容：

**图 4. 数据仓库建模阶段划分**



从上图我们可以清楚地看出，数据仓库的数据建模大致分为四个阶段：

1.   **业务建模**，这部分建模工作，主要包含以下几个部分：

* 划分整个单位的业务，一般按照业务部门的划分，进行各个部分之间业务工作的界定，理清各业务部门之间的关系。
* 深入了解各个业务部门的内具体业务流程并将其程序化。
* 提出修改和改进业务部门工作流程的方法并程序化。
* 数据建模的范围界定，整个数据仓库项目的目标和阶段划分。

2.   **领域概念建模**，这部分得建模工作，主要包含以下几个部分：

* 抽取关键业务概念，并将之抽象化。
* 将业务概念分组，按照业务主线聚合类似的分组概念。
* 细化分组概念，理清分组概念内的业务流程并抽象化。
* 理清分组概念之间的关联，形成完整的领域概念模型。

3.   **逻辑建模**，这部分的建模工作，主要包含以下几个部分：

* 业务概念实体化，并考虑其具体的属性
* 事件实体化，并考虑其属性内容
* 说明实体化，并考虑其属性内容

4.   **物理建模**，这部分得建模工作，主要包含以下几个部分：

* 针对特定物理化平台，做出相应的技术调整
* 针对模型的性能考虑，对特定平台作出相应的调整
* 针对管理的需要，结合特定的平台，做出相应的调整
* 生成最后的执行脚本，并完善之。

从我们上面对数据仓库的数据建模阶段的各个阶段的划分，我们能够了解到整个数据仓库建模的主要工作和工作量，希望能够对我们在实际的项目建设能够有所帮助。

### 3.4 数据仓库建模方法

大千世界，表面看五彩缤纷，实质上，万物都遵循其自有的法则。数据仓库的建模方法同样也有很多种，每一种建模方法其实代表了哲学上的一个观点，代表了一种归纳，概括世界的一种方法。目前业界较为流行的数据仓库的建模方法非常多，这里主要介绍范式建模法，维度建模法，实体建模法等几种方法，每种方法其实从本质上讲就是从不同的角度看我们业务中的问题，不管从技术层面还是业务层面，其实代表的是哲学上的一种世界观。我们下面给大家详细介绍一下这些建模方法。

**1. 范式建模法（Third Normal Form，3NF）**

范式建模法其实是我们在构建数据模型常用的一个方法，该方法的主要由 Inmon 所提倡，主要解决关系型数据库得数据存储，利用的一种技术层面上的方法。目前，我们在关系型数据库中的建模方法，大部分采用的是三范式建模法。

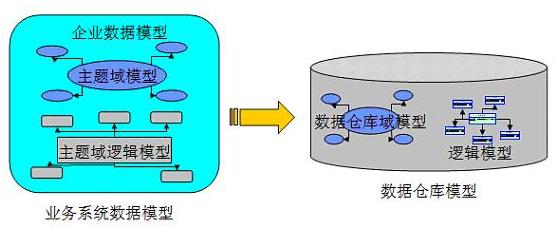
范式是数据库逻辑模型设计的基本理论，一个关系模型可以从第一范式到第五范式进行无损分解，这个过程也可称为规范化。在数据仓库的模型设计中目前一般采用第三范式，它有着严格的数学定义。从其表达的含义来看，**一个符合第三范式的关系必须具有以下三个条件 :**

* 每个属性值唯一，不具有多义性 ;
* 每个非主属性必须完全依赖于整个主键，而非主键的一部分 ;
* 每个非主属性不能依赖于其他关系中的属性，因为这样的话，这种属性应该归到其他关系中去。

由于范式是基于整个关系型数据库的理论基础之上发展而来的，因此，本人在这里不多做介绍，有兴趣的读者可以通过阅读相应的材料来获得这方面的知识。

根据 Inmon 的观点，数据仓库模型得建设方法和业务系统的企业数据模型类似。在业务系统中，企业数据模型决定了数据的来源，而企业数据模型也分为两个层次，即主题域模型和逻辑模型。同样，主题域模型可以看成是业务模型的概念模型，而逻辑模型则是域模型在关系型数据库上的实例。

**图 5. 范式建模法**



从业务数据模型转向数据仓库模型时，同样也需要有数据仓库的域模型，即概念模型，同时也存在域模型的逻辑模型。这里，**业务模型中的数据模型和数据仓库的模型稍微有一些不同。主要区别在于：**

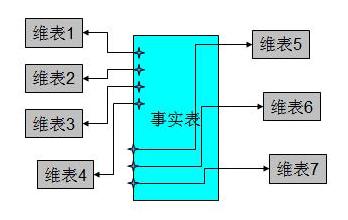
* 数据仓库的域模型应该包含企业数据模型的域模型之间的关系，以及各主题域定义。数据仓库的域模型的概念应该比业务系统的主题域模型范围更加广。
* 在数据仓库的逻辑模型需要从业务系统的数据模型中的逻辑模型中抽象实体，实体的属性，实体的子类，以及实体的关系等。

以笔者的观点来看，Inmon 的范式建模法的最大优点就是从关系型数据库的角度出发，结合了业务系统的数据模型，能够比较方便的实现数据仓库的建模。但其缺点也是明显的，由于建模方法限定在关系型数据库之上，在某些时候反而限制了整个数据仓库模型的灵活性，性能等，特别是考虑到数据仓库的底层数据向数据集市的数据进行汇总时，需要进行一定的变通才能满足相应的需求。因此，笔者建议读者们在实际的使用中，参考使用这一建模方式。

**2. 维度建模法**

维度建模法，Kimball 最先提出这一概念。其最简单的描述就是，按照事实表，维表来构建数据仓库，数据集市。这种方法的最被人广泛知晓的名字就是星型模式（Star-schema）。

**图 6. 维度建模法**



上图的这个架构中是典型的星型架构。星型模式之所以广泛被使用，在于针对各个维作了大量的预处理，如按照维进行预先的统计、分类、排序等。通过这些预处理，能够极大的提升数据仓库的处理能力。特别是**针对 3NF 的建模方法，星型模式在性能上占据明显的优势。**

同时，维度建模法的另外一个优点是，维度建模非常直观，紧紧围绕着业务模型，可以直观的反映出业务模型中的业务问题。不需要经过特别的抽象处理，即可以完成维度建模。这一点也是维度建模的优势。

但是，维度建模法的缺点也是非常明显的，由于在构建星型模式之前需要进行大量的数据预处理，因此会导致大量的数据处理工作。而且，当业务发生变化，需要重新进行维度的定义时，往往需要重新进行维度数据的预处理。而在这些与处理过程中，往往会导致大量的数据冗余。

另外一个维度建模法的缺点就是，如果只是依靠单纯的维度建模，不能保证数据来源的一致性和准确性，而且在数据仓库的底层，不是特别适用于维度建模的方法。

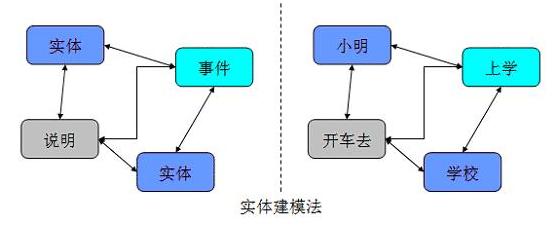
因此以笔者的观点看，维度建模的领域主要适用与数据集市层，它的最大的作用其实是为了解决数据仓库建模中的性能问题。维度建模很难能够提供一个完整地描述真实业务实体之间的复杂关系的抽象方法。

**3. 实体建模法**

实体建模法并不是数据仓库建模中常见的一个方法，它来源于哲学的一个流派。从哲学的意义上说，客观世界应该是可以细分的，客观世界应该可以分成由一个个实体，以及实体与实体之间的关系组成。那么我们在数据仓库的建模过程中完全可以引入这个抽象的方法，将整个业务也可以划分成一个个的实体，而每个实体之间的关系，以及针对这些关系的说明就是我们数据建模需要做的工作。

虽然实体法粗看起来好像有一些抽象，其实理解起来很容易。即我们可以将任何一个业务过程划分成 3 个部分，实体，事件和说明，如下图所示：

**图 7. 实体建模法**



上图表述的是一个抽象的含义，如果我们描述一个简单的事实：“小明开车去学校上学”。以这个业务事实为例，我们可以把“小明”，“学校”看成是一个实体，“上学”描述的是一个业务过程，我们在这里可以抽象为一个具体“事件”，而“开车去”则可以看成是事件“上学”的一个说明。

从上面的举例我们可以了解，我们使用的抽象归纳方法其实很简单，**任何业务可以看成 3 个部分：**

* **实体**，主要指领域模型中特定的概念主体，指发生业务关系的对象。
* **事件**，主要指概念主体之间完成一次业务流程的过程，特指特定的业务过程。
* **说明**，主要是针对实体和事件的特殊说明。

由于实体建模法，能够很轻松的实现业务模型的划分，因此，在业务建模阶段和领域概念建模阶段，实体建模法有着广泛的应用。从笔者的经验来看，再没有现成的行业模型的情况下，我们可以采用实体建模的方法，和客户一起理清整个业务的模型，进行领域概念模型的划分，抽象出具体的业务概念，结合客户的使用特点，完全可以创建出一个符合自己需要的数据仓库模型来。

但是，实体建模法也有着自己先天的缺陷，由于实体说明法只是一种抽象客观世界的方法，因此，注定了该建模方法只能局限在业务建模和领域概念建模阶段。因此，到了逻辑建模阶段和物理建模阶段，则是范式建模和维度建模发挥长处的阶段。

因此，笔者建议读者在创建自己的数据仓库模型的时候，可以参考使用上述的三种数据仓库得建模方法，在各个不同阶段采用不同的方法，从而能够保证整个数据仓库建模的质量。

# 数据仓库建模案例

### O2O行业案例分析

#### 4.1.1订单表，订单商品表，用户表建模