《数据仓库工具箱——维度建模权威指南》笔记

AUTHOR: 彭玲 TIME: 2023/3/31

抵制规范化的冲动

---

1. 数据仓库、商业智能及维度建模初步

数据获取与数据分析的区别

- 操作型系统保存数据
  - 不维护历史数据，仅保留最新的系统状态
  - 强事务性，更快的处理事务，一次处理一个事务
- 分析型系统（DW/BI）使用数据
  - 维护历史数据，且数据量大
  - 一次处理多个事务

数据仓库与商业智能的目标

- 方便地存取信息
- 以一致的形式展现信息
- 必须能适应变化
- 必须能及时展现信息
- 必须成为保护信息财富的安全堡垒（有效控制对组织中机密信息的访问）
- 必须成为提高决策制定能力的权威和可信的基础
- 系统成功的标志是业务群体接收 DW/BI 系统

数据仓库和商业智能的成功需要一整套的技能，包括数据库管理的技能，也包括商业分析师的技能。

维度建模简介

维度建模需要同时满足2个需求：

- 以商业用户可理解的方式发布数据
- 提供高效的查询性能（高性能检索）

从简单的数据模型开始是保持设计简单性的基础。复杂的数据模型会导致查询性能低下，最终使商业用户反感。

维度模型可以不满足第3范式（3NF，规范化模型），3NF 用于消除冗余但不利于理解查询。

维度建模包含的信息与规范化模型包含的信息相同，但将数据以一种用户可理解的、满足查询性能要求的、灵活多变的方式进行了包装。

星型模式与 OLAP 多维数据库

星型模式：关系数据库管理系统中实现的维度模型。

OLAP（OnLine Analytical Processing，联机分析处理）：在多维数据库环境中实现的维度模型。

OLAP 部署注意事项：

- 构建于关系数据库之上的星型模式是建立 OLAP 多维数据库的良好物理基础

- 相较于 OLAP 多维数据库，随着计算机硬件与软件的发展，RDBMS 性能优势不突出
- OLAP 多维数据库的部署细节与选择的提供商有关
- OLAP 多维数据库提供更多的复杂安全选项
- OLAP 多维数据库能够提供更加丰富的分析能力，可以作为选择 OLAP 产品的主要依据
- 当需要使用其他缓慢变化维度技术重写数据时，OLAP 多维数据库通常需要被全部或部分地重新处理
- OLAP 多维数据库方便地支持事务和周期性快照事实表，但无法处理累积快照事实表
- OLAP 多维数据库支持具有层次不确定的复杂的不规则层次结构
- OLAP 多维数据库与关系数据库比较，能对实现下钻层次的维度关键词结构提供更详细的约束
- 一些 OLAP 产品无法确保实现维度角色和别名，因此需要定义不同的物理维度

用于度量的事实表

维度模型中的事实表（“事实”这一术语表示某个业务度量）存储组织机构业务过程事件的性能度量结果。

维度建模的核心原则之一：同一事实表中的所有度量行必须具有相同的粒度（每行中的数据是一个特定级别的细节数据，称为“粒度”）。

维度建模的基本原则：物理世界的每一个度量事件与对应的事实表行具有一对一的关系。

事实表存在稀疏特性。

事实表在行数上趋向于变长，在列上趋向于变短。

事实表的粒度可以划分为三类：事务（最常见）、周期性快照和累积快照。

事实表表示多对多关系，其他表称为维度表。事实表的主键常称为组合键。

用于描述环境的维度表

维度表包含与业务过程度量事件有关的文本环境。

与事实表相比，维度表趋向于包含较少的行，但由于可能存在大量文本列而导致存在多列的情况。

维度表的属性是所有查询约束和报表标识的来源。属性应该包含真实使用的词汇而不是令人感到迷惑的缩写。

区分数值属性是事实属性还是维度属性的方法：

- 如果一个数值属性包含多个值并作为计算的参与者的度量——事实属性
- 如果仅是对具体值的描述，是一个常量、某一约束和行标识的参与者——维度属性

维度表不需要满足 3NF，它常常是非规范化的。

星型模式中维度与事实的连接

维度模型表示每个业务过程包含事实表，事实表存储事件的数值化度量、多个维度外键，围绕事实表的是多个维度表。这种类似星状的结构通常称为“星形连接”。

Kimball 的 DW/BI 架构

共有四个组成部分：操作型源系统、ETL 系统、数据展现和商业智能应用。

操作型源系统

- 处于数据仓库之外，用于获取业务事务
- 一般不维护历史信息

ETL 系统

ETL (Extract Transformation and Load, 获取-转换-加载) 系统包括：一个工作区间、实例化的数据结构以及一个过程集合。

ETL 系统是处于操作型源系统与 DW/BI 展现系统之间的区域。

- E：读取并理解源数据，并将需要的数据复制到 ETL 系统中
- T：多种转换操作，例如，清洗数据（消除拼写错误、解决领域冲突、处理错误的元素、解析为标准格式），合并不同数据源的数据，复制数据等
- L：加载数据到展现区域

不能在用户查询中使用规范化结构，因为规范化结构难以同时满足可理解性和性能这两个目标。

用于支持商业智能决策的展现区

DW/BI 展现区用于组织、存储数据，支持用户、报表制作者以及其他分析型 BI 应用的查询。

- 数据应该以维度模型（采用星型模式或者 OLAP 多维数据库）来展现
- 必须包含详细的原子数据（最细粒度的数据）
- 展现区的数据可以围绕业务过程度量事件来构建
- 必须使用公共的、一致性的维度建立维度结构（遵守总线结构）

处于 DW/BI 系统的可查询展现区中的数据必须是维度化的、原子的、以业务过程为中心的。

商业智能应用

“BI 应用”这一术语泛指为商业用户提供利用展现区制定分析决策的能力（所有 BI 应用的查询针对的是 DW/BI 展现区）。

其他 DW/BI 架构

独立数据集市架构

按部门独立建设，不需要考虑跨组织的数据控制和协调问题。（不提倡使用）

辐射状企业信息工厂 Inmon 架构

辐射状企业信息工厂（Corporate Information Factory, CIF）获得的原子数据保存在满足 3NF 的数据库中，这种规范化的、原子数据的仓库被称为 CIF 架构下的企业数据仓库（Enterprise Data Warehouse, EDW）。规范化的 EDW 是 CIF 中强制性的构件。

- CIF 利用规范化的 EDW
- Kimball 架构强调具有一致性维度的企业总线的重要作用

维度建模神话

常见的误解：

- 维度模型仅包含汇总数据（应该包含细节数据）
- 维度模型是部门级而不是企业级（应该围绕业务过程组织而不是部门）
- 维度模型是不可扩展的（实际上非常易于扩展）
- 维度模型仅用于预测（应该以度量过程为中心、适应变化的）
- 维度模型不能被集成（遵守企业数据仓库总线结构，维度模型多数都能被集成）

2. Kimball 维度建模技术概述

基本概念

- 收集业务需求与数据实现
  - 与业务代表交流发现需求
  - 与源系统专家交流数据实际情况
- 维度设计过程（4步骤）
  - 选择业务过程
  - 声明粒度
  - 确认维度
  - 确认事实
- 粒度：用于确定某一事实表中的行表示什么。

  选择维度或事实前必须声明粒度，因为每个候选维度或事实必须与定义的粒度保持一致。
- 描述环境的维度

  维度提供围绕某一业务过程事件所涉及的"谁、什么、何处、何时、为什么、如何"等背景。
- 用于度量的事实

  事实涉及来自业务过程事件的度量，基本上都是以数量值表示。

事实表技术基础

- 事实表结构

  从最低的粒度级别来看，事实表行对应一个度量事件。事实表的设计完全依赖于物理活动。事实表存储：
  - 现实世界的操作型事件产生的可度量数值
  - 相关维度表的外键
  - 可选的退化维度键和日期/时间戳
- 可加、半可加、不可加事实

  事实表中的数字维度可划分为三类：
  - 可加：可以按照事实表中的任意维度汇总
  - 半可加：可以对事实表中的某些维度汇总（不能对所有维度汇总）
  - 不可加：如比率
- 事实表中的空值

  可以存在空值度量，但外键不能存在空值。
- 一致性事实

  如果某些度量出现在不同的事实表中，应保证对事实的技术定义相同（具有相同的命名），如果它们不兼容，应该以不同的命名告诫业务用户和 BI 应用。
- 事务事实表

  事务事实表的一行对应空间或时间上的某点（事务）的度量事件，包含一个与维度表关联的外键。
- 周期快照事实表

  每一行对应标准周期（某天、某周、某月）的度量事件。粒度是周期性的，而不是个体的事务。

- 累积快照事实表

  累积快照事实表的行汇总了发生在过程开始和结束之间可预测步骤内的度量事件。

- 无事实的事实表

  没有度量（可记录的数字化事实），仅有关系（外键）的事实表。

- 聚集事实表或 OLAP 多维数据库

  聚集事实表是对原子粒度事实表数据进行简单的数字化上卷操作，目的是为了提高查询性能。

  聚集事实的构建是通过汇总多个原子事实表的度量获得的。

- 合并事实表

  为了能够带来方便，将来自多个过程的、以相同粒度表示的事实合并为一个单一的合并事实表。

  合并事实表会增加 ETL 处理过程的负担，但降低了 BI 应用的分析代价。

维度表技术基础

- 维度表结构

  每个维度表都包含单一的主键列。

  维度表通常较宽，是扁平型非规范表。

  维度表属性是查询及 BI 应用的约束和分组定义的主要目标。

- 维度代理键

  一般维度表的主键是一个自增 id（维度代理键），日期维度的维度表不需要遵循代理键规则。

- 自然键、持久键和超自然键

  自然键例子：工号

  超自然键或持久键例子：离职员工重新入职，改键不会发生变化的键

- 维度表中的空值属性

  推荐采用描述性字符串替代空值。例如，使用 Unknown 或 Not Applicable 替换空值。

  避免在维度属性中使用空值，因为不同的数据库系统在处理分组和约束时，针对空值的处理方法不一致。

- 扮演角色的维度

  在维度建模中，角色是指在一个事实表（Fact Table）中，一个维度（Dimension）扮演的不同角色。例如，在一个销售事实表中，时间维度可以扮演多个角色，如销售日期、发货日期、订单日期等。

- 雪花维度

  维度表按照层次结构建立，称之为雪花模式。雪花模式查询困难、查询性能低，应该避免使用。应该使用扁平化的非规范的维度表。

- 支架维度

  某个维度表引用其他维度表，被引用的辅助维度称为支架维度。少用支架维度，由事实表实现维度之间的关联。

使用一致性维度集成

- 一致性维度

  当不同的维度表的属性具有相同的列名和领域内容时，称维度表具有一致性。

- 价值链

用于区分组织中主要业务过程的自然流程。例如，销售商的价值链可能包括购买、库存、零售额等。

- 企业数据仓库总线矩阵

  是用于设计并与企业数据仓库总线架构交互的基本工具。

  矩阵的行表示业务过程，列表示维度。矩阵中的点表示维度与给定的业务过程是否存在关联关系。

处理缓慢变化维度属性

- 原样保留：维度属性值不会发生变化。
- 重写：维度行中原来的属性被新值覆盖。
- 增加新行
- 增加新属性
- 增加微型维度

高级事实表技术

- 事实表代理键

  代理键可用作所有维度表的主键。

高级维度技术

- 多值维度

  多值维度通常用于描述具有多个属性或特征的数据。在数据模型中，多值维度可以被看作是一个具有多个子维度的维度。例如，一个人可以有多个爱好，每个爱好可以被看作是一个子维度，而人这个维度就是一个多值维度。

- 步骤维度

  步骤维度（Step Dimension）是一种用于描述业务过程中各个步骤的维度，通常用于数据仓库和商业智能中。

  步骤维度通常包括以下几个步骤：

  1. 定义业务过程：首先需要定义业务过程中的各个步骤，例如订单处理过程、客户服务过程等。
  2. 确定步骤指标：对于每个业务过程步骤，需要确定用于衡量其性能的指标。例如，对于订单处理过程，可以使用订单处理时间、订单处理成功率等指标来衡量性能。
  3. 设计步骤维度表：在设计步骤维度时，需要为每个业务过程步骤创建一个维度成员，并为每个成员分配一个唯一的标识符。同时，需要将每个步骤的指标值存储在维度表中。
  4. 关联事实表：将步骤维度表与事实表进行关联，以便可以对业务过程进行分析。在事实表中，可以使用步骤维度表中的标识符来识别每个业务过程步骤，并使用步骤维度表中的指标来衡量性能。

  通过使用步骤维度，可以深入了解业务过程中各个步骤的性能，并确定哪些步骤需要优化。同时，步骤维度还可以帮助我们跟踪业务过程中各个步骤的历史性能，并预测未来的趋势。

3. 零售业务

维度模型设计的4步过程

第 1 步：选择业务过程

业务过程包含以下公共特征：

- 业务过程一般是一个行为动词，表示业务执行过程中的活动。
- 业务过程通常由操作型系统支撑，例如，账单或购买系统。

- 业务过程建立或获取关键性能度量。
- 业务过程通常由输入激活，产生输出度量。

第 2 步：声明粒度

声明粒度意味着精确定义某个事实表的每一行表示什么。以业务术语表示粒度，典型的粒度声明如下：

- 客户销售事务上的每个产品扫描到一行中
- 医生开具的票据的列表内容项采用一行表示
- 机场登机口处理的每个登机牌采用一行表示
- 每个银行账户每月的情况采用一行表示

第 3 步：确定维度

维度要解决的是：业务人员如何描述来自业务过程度量事件的数据？

常见的维度实例包括：日期、产品、客户、雇员、设备等。

第 4 步：确定事实

事实的确定：回答 “过程的度量是什么？” 这一问题。

典型的事实是可加性数值，例如，订货数量或成本总额等。

零售业务案例研究

第 1 步：选择业务过程

通过对业务需求以及可用数据源的综合考虑，选择 POS 零售交易。

第 2 步：声明粒度

粒度较高的模型无法实现用户下钻细节的需求。DW/BI 系统要求数据尽可能最细粒度表示。

最细粒度的数据是 POS 交易的单个产品。

第 3 步：确定维度

描述性维度：日期、产品、门店、促销、收银员、支付方式。

第 4 步：确定事实

事实必须与粒度吻合：放入 POS 交易的单独产品线项。

POS 系统收集的事实包括：销售数量、单价、折扣、净支付价格、扩展折扣、美元销售额等。

维度表设计细节

日期维度

- 假日标识：有两个可能值，用有意义的值（例如 `假日` 或 `非假日`，而不是 `Y/N`、`1/0`、`真/假`）表示。
- 当前与相对日期属性：需要更新。

产品维度

- 需要扁平化多对一层次（即使大量值是重复的）。规范化这些值（放入不同的表）将难以简单化和高性能。

- 作为属性或事实的数字值：需要判断是否要放入事实表中。
    - 数字值用于计算目的或者变化分析，则可能应该属于事实表。
    - 数字值（稳定的）用于过滤和分组，则应该属于维度属性。
    - 数字值同时用于计算以及过滤/分组，则应该在事实表和维度表中同时存储该值。
- 下钻维度属性
    - 下钻：从维度表中增加行头指针属性。
    - 上卷：从维度表中移除行表头。

商店维度

商店维度描述零售连锁店的每个门店。

促销维度

促销维度描述了销售商品的促销条件（包括：临时降价、终端通道展示、报纸广告、礼券等）。

关于空值：

- 不要在事实表中使用空值键。
- 建议用描述性字符串（`Unknown`或`Not Applicable`等）替换空值。

事务号码的退化维度

退化维度（Degenerate Dimension, DD）就是那些看起来像是事实表的一个维度关键字，但实际上并没有对应的维度表。

零售模式的扩展能力

过早地聚集和汇总限制了增加补充维度的能力，因为增加的维度通常无法在更高粒度级别上应用。

针对各种变化，如何处理？

- 新维度属性：维度表添加新列
- 新维度：添加新的维度表，向事实表中添加新的外键列
- 新可度量事实：向事实表中添加数据（同一粒度），创建新的事实表（不同粒度）

维度与事实表键

- 维度表代理键的作用：连接维度表与事实表。

  数据仓库中维度表与事实表的每个连接应该基于无实际含义的整数代理键，应该避免使用自然键作为维度表的主键。

- 自然键通常被建模为维度表的属性。
- 超自然键：永久的持久性标识符。

- 日期维度：具有可预测性。日期维度的主键是一个有意义的整数，格式为`YYYYMMDD`，常用于分区。
- 事实表中推荐但不强制使用代理键（是一个简单整数）。

抵制规范化的冲动

- 具有规范化维度的雪花模式（规范化的维度表被称为`雪花模式`），抵制原因：易用性和性能。
    - 简单化是维度建模的主要目标之一

    - 雪花模式下，大量的表和连接操作常导致查询性能缓慢
    - 雪花模式下，磁盘空间节省不明显（牺牲一些维度空间有利于改善性能和可用性）
    - 雪花模式无法实现位图索引（位图索引能提高查询或单一列约束的性能）
- 蜈蚣事实表，结构类似于一直蜈蚣，具有一个中央的事实表，周围围绕着多个维度表，形成多个分支。
    - 大量的连接对可用性和性能都是问题
    - 多数业务过程，事实表关联的维度不超过20个（如果超过25个，应采取措施合并关联的维度）