维度建模技术实践——深入事实表

原创 云祁 云祁的数据江湖 2020-12-06 13:36

收录于合集
#进击的大数据 41 #数据建模 11

大家好，我是云祁！

前面我们聊过了维度建模的灵魂所在——维度表设计，今天就深入学习下维度建模的核心——事实表。

聊聊维度建模的灵魂所在——维度表设计

事实表是维度建模的核心表和基本表。

它存储了业务过程中的各种度量和事实，而这些度量和事实正是下游数据使用人员所要关心和分析的对象。

目前事实表主要探讨三种：

- 事务事实表
- 快照事实表
- 累计快照事实表

还有一种特殊的事实表——无事实的事实表，最后还将讨论事实表的聚集和汇总。

事务事实表

事务事实表是维度建模事实表中最为常见、使用最为广泛的事实表。

事务事实表通常用于记录业务过程的事件，而且是原子粒度的事件。事务事实表中的数据在事务事件发生之后，数据的粒度通常是每个事务一条记录。一旦事务被提交，事实表数据被插入，数据就不再进行更改。

我们通过事务事实表存储单次业务事件 / 行为的细节，以及存储与事件相关的维度细节，用户即可以单独或者聚合分析业务事件和行为。

事务事实表的粒度确定是事务事实表设计过程中的关键步骤，一般都会包含可加的度量和事实。理解概念的最佳途径无疑是实际的例子，因此下面将结合超市零售业务以及维度建模的四个环节来说明事务事实表。

（1）选择业务过程

在超市的零售示例中，业务用户做的事情是更好的理解POS系统记录顾客购买的情况，那么很容易确定业务过程就是POS系统记录的顾客购买情况，即在什么时候、什么商品、哪个收银台、销售了哪些产品等。

（2）定义粒度

顾客单次购买行为的体现是一张购物小票，但是事务事实表应该选择最原子粒度的事件，所以小票的子项（在业务上的动作即为收银员每次扫描的商品条码）应该是超市零售事务事实表的粒度。

（3）确定维度

小票子项的粒度确定后，销售日期、销售商品、销售收银台、销售门店等维度很容易被确定了。另一个不太容易考虑到的是维度是促销行为，但是通过和业务人员交流或者查看报表表头等也能够发现此维度。

（4）确定事实

维度设计的最后一步，是确定哪些事实和度量应该在事实表中出现。对于本例，商品销售数量、销售价格和销售金额很容易确定下来。但是实际上，商品的成本价是确定的，因此可以很容易地确定商品的销售毛利：（商品实际销售价格－商品成本价） 销售数量，基于下游使用便利这一因素，也应该将此放人事务事实表中。

基于毛利润也可以计算出毛利率，那么毛利率这种比例应该放入事务事实表吗？在事实表的设计中，一个常见的原则是只存放比例的分子和分母，因此比例的计算是和业务强，业务逻辑可能非常的复杂，所以一般不加入事实表中。

至此，我们也完成了超市零售事务的事实表和维度表的设计，超市零售事务事实表以及相关的维度表如图所示：

快照事实表

在实际的业务活动中，除了关心单次的业务事件和行为外，很多时候还关心业务的状态（当前状态、历史状态）。以超市零售业务为例，管理人员和分析人员除了关心销售情况，还会关心商品的库存情况，例如哪些商品的库存情况，例如哪些商品库存告罄需要补货、哪些积压需要促销，而这正是**快照事实表（也叫周期快照事实表）**所要解决发范畴。

所谓周期快照事实表，是指间隔一定的周期对业务的状态进行一次拍照并记录下来的事实表。最常见的例子是销售库存、银行账户余额等。

与事务事实表的稀疏性不同（这里的稀疏性是相对的），周期快照事实表通常被认为是稠密的。因此事务事实表只有事务发生才会记录，但是周期快照则必须捕获当前每个实体的状态。

比如，某个商品如果某天没有销售，那么这个商品不会存在于当天的事务事实表中的，但是为了记录其库存情况，即使没有销售行为，也必须再周期快照事实表中对其进行拍照。

周期快照事实表的周期通常需要和业务方共同确定，最常见的周期是天、周和月等。

周期快照事实表中的事实一般是半可加的，如某个商品的库存可以跨商品、仓库等相加，但是明显在时间上相加是没有意义的。

下面就以超市的库存业务为例来介绍周期快照事实表的设计过程。

（1）选择业务过程

本例是为了更好地理解超市的库存情况，因此业务过程就是商品的库存情况，即在什么时候、什么商品、哪个仓库的库存量如何。

（2）定义粒度

这里的粒度主要指库存的周期，商品的粒度很容易确定（注意这里是 SKU 级别）。选择库存的周期需要考虑到数据量膨胀情况。

考虑如下例子，某个超市有 万个商品（即SKU）， 其有 100 家连锁店，那么每天对其库存拍照将有 100*10000=100 万行记录，那么一年将有 365*1000000=3.65亿条记录。当然随着目前存储的日益廉价，这些都不是问题，但是设计人员需要考虑到这些因素。

（3）确认维度

对于超市零售库存，相应的维度为周期（天 周、月等） 商品、仓库（总仓、分仓或者门店等）。

（4）确定事实

这里的事实很容易确定，即库存量。但是仅仅记录现存库存是不够充分的，因为业务上通常会和其他事实协同来度量库存的变化趋势、快慢等，所以还可对周期快照事实表的事实进行增强 。

基于上述设计的周期快照事实表及相关维度如图所示：

累计快照事实表

事实表的第三种类型是累计快照事实表，相比前两者，累计快照事实表没那么常见，但是对于某些业务场景来说非常有价值。

累计快照事实表非常适用于具有工作流或者流水线形式业务的分析，这些业务通常涉及多个时间节点或者有主要的里程碑事件，而累计快照事实表正是从全流程角度对其业务状态的拍照。

考虑车险理赔业务，一次车险的理赔通常包括客户报案、保险公司立案、客户提交理赔材料、理赔审批通过和付款等关键步骤，而累积快照事实表正是从全流程角度对每个车险理赔单的拍照，拍照内容

即是其关键步骤的各个状态，便于业务人员一目了然地分析各个理赔单的状态、步骤间的耗时等。

下面以车险理赔业务为例来介绍累计周期快照事实表。

（1）选择业务过程

本例是为了更好地理解保险公司的车险理赔业务，因此业务过程就是车险理赔，即在什么时候、 哪个理赔申请所处的状态如何。

（2）定义粒度

累计周期快照事实表的粒度一般很容易确定，就是业务的某个实体，这里即为保险理赔申请。

（3）确定维度

对于累计周期快照事实表，相关的维度包含快照周期（天、周、月 和年等）、理赔申请人、受理 、审核人、网点 电话或者实体）等。

（4）确定事实

这里的事实包括索赔金额、审批金额、打款金额、处理时长等。

无事实的事实表

在维度建模中，事实表是过程度量的核心，也是存储度量的地方 但事实表并不总是需要包含度量和事实，这类不包含事实的事实表被称为 **无事实的事实表**。

乍一听有点奇怪，但是请考虑下面业务场景，银行客户服务中心接受客户电话咨询或者在线业务咨询，这里并没有任务的业务度量值，唯一的度量值就是单次咨询事件。其他类似事件还有学生课程出席情况、用户在网站上的浏览行为、客户对广告的点击行为等。

无事实的事实表通常人为增加一个常量列（其列的值总是为 1) 来方便对业务时间的统计分析。

以学生在各门课程中的出席情况为例给出无事实的事实表的维度设计方案：

总结

在经典的维度建模事实表设计中，事实表将仅存储维度表外键、选定的度量以及退化维度等，例如我们前面提到的超市零售事务事实表。

这样的设计主要是出于存储的成本以及处理的性能考虑 ，如果把维度的属性字段等都放在事实表中，那么将带来大量的存储开销，而且处理性能也将大大受到影响。但是在大数据时代，随着 HDFS、MapReduce 为代表的各种分布式存储和计算技术的发展，存储成本以及性能等不再是关键，所以在维度建模理论反规范化思想的基础上，可以更进一步地把常用的维度属性沉淀在事实表中，这样下游使用更为直接和便捷，不需要每次都关联相关维度表获取有关维度属性 也就是说，以存储的冗余为代价，换来了下游的使用便捷以及多次关联计算开销，而在大数据时代，这是完全划算的。

浅谈大数据建模的关键技术：维度建模