详解维度建模之事实表

每个数据仓库都包含一个或者多个事实数据表。其中可能包含业务销售数据，如现金登记事务所产生的数据，通常包含大量的行。事实数据表的主要特点是包含数字数据（事实），并且这些数字信息可以汇总，以提供有关单位作为历史的数据，每个事实数据表包含一个由多个部分组成的索引，该索引包含作为外键的相关性维度表的主键，而维度表包含事实记录的特性。有很多规范大家可以参考：数据仓库建设规范（文档版），严选-数仓规范和评价体系

— 01 —

事实表基础

事实表特征

事实表作为数仓维度建模的核心，紧紧围绕着**业务过程**来设计，通过获取描述业务过程的度量来表达业务过程，包含了引用的维度和业务过程有关的度量。事实表中一条记录所表达的业务细节程度被称为**粒度**(业务中的细节粒度)。通常粒度可以通过两种方式来表达：一种是维度属性组合所表示的细节程度，另一种是所表示的具体业务含义。

作为**度量业务过程**的事实（事实表属性），一般为整型或浮点型的十进制数值，有可加性、半可加性和不可加性三种类型：

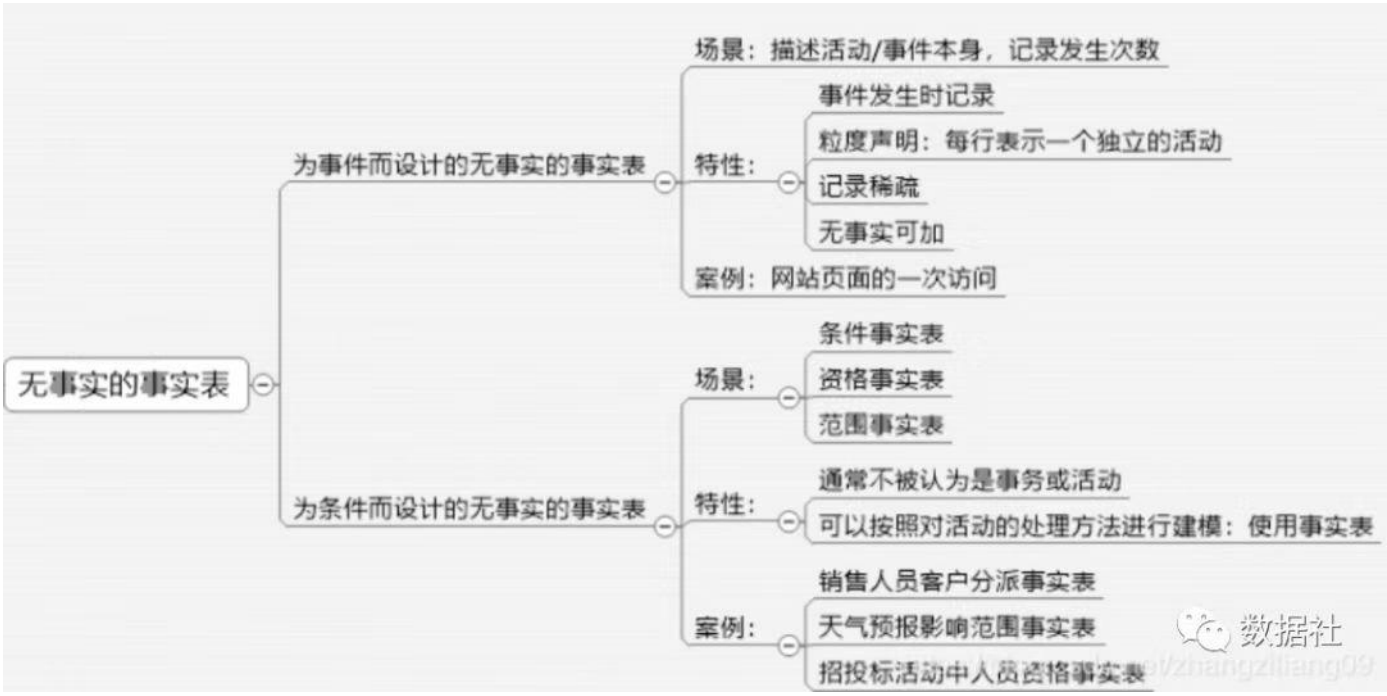
- **可加性事实** 是指可以按照与事实表关联的任意维度进行汇总。
- **半可加性事实** 只能按照特定维度汇总，不能对所有维度汇总，比如库存可以按照地点和商品进行汇总，而按时第门章事实表设计丁一间维度把一年中每个月的库存累 加起来则毫无意义。
- **不可加事实** 不具备可加性，比如比率型事实。对于不可加性事实可分解为可加的组件来实现聚集。

有事实的事实表

有事实表分为三种类型： **事务事实表、周期快照事实表**和**累积快照事实表**。

无事实的事实表

无事实的事实表可以用来跟踪事件的发生。例如，在给定的某一天中发生的学生参加课程的事件，可能没有可记录的数字化事实，但该事实行带有一个包含日期、学生、教师、地点、课程等定义良好的外键。利用无事实的事实表可以按各种维度计数上课这个事件。



— 02 —

事实表设计规则

- **尽可能包含所有与业务过程相关的事实；**
- **只选择与业务过程相关的事实；**
- **分解不可加性事实为可加的组件；**比如订单的优惠率，应该分解为订单原价金额与订单优惠金额
- **在选择维度和事实之前必须先声明粒度；**
- **在同一个事实表中不能有多种不同粒度的事实；**粒度的声明是事实表设计中不可忽视的重要一步，粒度用于确定事实表中一行所表示业务的细节层次，决定了维度模型的扩展性，在选择维度和事实之前必须先声明粒度，且每个维度和事实必须与所定义的粒度保持一致
- **在同一个事实表中不能有多种不同粒度的事实；**
- **事实的单位要保持一致；**
- **对事实的 null 值要处理；**在数据库中null值对常用的大于或小于等SQL不生效，建议使用零值填充
- **使用退化维度提高事实表的易用性；**目的主要是为了减少下游用户使用时关联多个表的操作。直接通过退化维度实现对事实表的过滤查询、控制聚合层次、排序数据以及定义主从关系等

事实表设计方法

Kimball的四步维度建模方法：**选择业务过程、声明粒度、确定维度、确定事实**。

Step 1:选择业务过程及确定事实表类型。

在明确了业务需求以后，接下来需要进行详细的需求分析，对业务的整个生命周期进行分析，明确关键的业务步骤，从而选择与需求有关的**业务过程**。(业务过程通常使用行为动词表示业务执行的活动)

Step 2:声明粒度 。

粒度的声明是事实表建模非常重要的一步，意味着精确定义事实表的每一行所表示的业务含义，粒度传递的是与事实表度量有关的细节层次。明确的粒度能确保对事实表中行的意思的理解不会产生混淆，保证所有的事实按照同样的细节层次记录。

Step 3 :确定维度 。

完成粒度声明以后，也就意味着确定了主键，对应的维度组合以及相关的维度字段就可以确定了，应该选择能够描述清楚业务过程所处的环境的维度信息。

Step 4 : 确定事实 。

事实可以通过回答“过程的度量是什么”来确定。应该选择与业务过程有关的所有事实，且事实的粒度要与所声明的事实表的粒度一致。事实有可加性、半可加性、非可加性三种类型 ， 需要将不可加性事实分解为可加的组件。

Step 5:冗余维度 。

冗余维度是在kimball维度建模方法基础上新增的步骤。主要是因为在大数据的事实表模型设计中，需要考虑更多的是提高下游用户的使用效率，降低数据获取的复杂性，减少关联的表数量。所以通常事实表中会冗余方便下游用户使用的常用维度，以实现对事实表的过滤查询、控制聚合层次、排序数据以及定义主从关系等操作。

— 04 —

有事实的事实表

有事实表分为三种类型： **事务事实表、周期快照事实表**和**累积快照事实表**。

事务事实表

单事务事实表，针对于**每个业务过程**设计一个事实表，方便每个业务过程进行独立分析研究。

优点：更方便跟踪业务流程细节数据，针对特殊的业务分析场景比较方便和灵活，数据处理上也更加灵活;

弊端：数仓中需要管理太多的事实表，同时跟踪业务流转不够直观

多事务事实表，将**不同的事实放到同一个事实表**中，即同一个事实表包含不同的业务过程。多事务事实表在设计时有两种方法进行事实的处理：

一是不同业务过程的事实使用不同的事实字段进行存放：

二是不同业务过程的事实使用同一个事实字段进行存放，但增加一个业务过程标签。

优点：能够更直观的跟踪业务流转和当前状态，流程事实集中，方便大部分的通用分析应用场景，由于和业务侧的数据模型设计思路一致，也是目前最常用的事实表设计;

弊端：细节数据跟踪不到位，特殊场景的分析不够灵活;

表 11.2　单事务事实表和多事务事实表的比较

	单事务事实表	多事务事实表
业务过程	一个	多个
粒度	相互间不相关	相同粒度
维度	相互间不相关	一致
事实	只取当前业务过程中的事实	保留多个业务过程中的事实，非当前业务过程中的事实需要置零处理
冗余维度	多个业务过程，则需要冗余多次	不同的业务过程只需要冗余一次
理解程度	易于理解，不会混淆	难以理解，需要通过标签来限定
计算存储成本	较多，每个业务过程都需要计算存储一次	较少，不同业务过程融合到一起，降低了存储计算量，但是非当前业务过程的度量存在大量零值

两种表的设计区别在于对业务流程的拆分思路不同，具体选择事实表的构建思路，需要根据实际的业务确定，一般建议两者结合。

父子事实的处理方式，通过分摊父订单的金额将所有业务过程的度量全部带进淘宝交易事务事实表中，包括下单数量、商品价格、子订单折扣、下单分摊比例、父订单支付金额、父订单支付邮费、父订单折扣、子订单下单金额、子订单下单有效金额、支付分摊比例、子订单支付金额等，将父子事实同时冗余到事务表中。

设计准则

1. **事实完整性**

   事实表包含与其描述的过程有关的所有事实，即尽可能多地获取所有的度量。

2. **事实一致性**

   在确定事务事实表的事实时，明确存储每一个事实以确保度量的一致性。

3. **事实可加性**

   事实表确定事实时，往往会遇到非可加性度量，比如分摊比例、利润率等，虽然它们也是下游分析的关键点，但往往在事务事实表中关注更多的是可加性事实，下游用户在聚合统计时更加方便。

周期快照事实表

快照事实表在确定的间隔内对实体的度量进行抽样，这样可以很容易地研究实体的度量值，而不需要聚集长期的事务历史。

特征

1. **用快照采样状态**

   快照事实表以预定的间隔采样状态度量。这种间隔联合一个或多个维度，将被用来定义快照事实表的粒度，每行都将包含记录所涉及状态的事实。

2. **快照粒度**

   事务事实表的粒度可以通过业务过程中所涉及的细节程度来描述，但快照事实表的粒度通常总是被多维声明，可以简单地理解为快照需要采样的周期以及什么将被采样。

3. **密度与稀疏性**

   快照事实表和事务事实表的一个关键区别在密度上。事务事实表是稀疏的，只有当天发生的业务过程，事实表才会记录该业务过程的事实，如下单、支付等；而快照事实表是稠密的，无论当天是否有业务过程发生，都会记录一行，比如针对卖家的历史至今的下单和支付金额，无论当天卖家是否有下单支付事实，都会给该卖家记录一行。

4. **半可加性**

在快照事实表中收集到的状态度量都是半可加的。与事务事实表的可加性事实不同，半可加性事实不能根据时间维度获得有意义的汇总结果。

设计实例

单维度的每天快照事实表

确定粒度、确定维度

混合维度的每天快照事实表

确定粒度、确定维度、确定状态度量

全量快照事实表

相比单维度的快照事实表，多了一些冗余维度。例如，商品评价表，多了子订单维度、商品维度、评论者维度。

累计快照事实表

对于类似于研究事件之间时间间隔的需求，采用累计快照事实表可以很好地解决。

如在统计买家下单到支付的时长、买家支付到卖家发货的时长等，事务事实表很难满足，需要用到累计快照事实表。

特征

1. **数据不断更新**

   针对于实体中的某一实例定期更新。

2. **多业务过程日期**

   此为累积快照事实表适用于具有较明确起止时间的短生命周期的实体，比如交易订单、物流订单等，对于实体的每一个实例，都会经历从诞生到消亡等一系列步骤。对于商品、用户等具有长生命周期的实体，一般采用周期快照事实表更合适。**累积快照事实表的典型特征是多业务过程日期，用于计算业务过程之间的时间间隔。**但结合阿里巴巴数据仓库模型建设的经验，对于累积快照事实表，还有一个重要作用是保存全量数据。

特殊处理

1. **非线性过程**

   淘宝一般流程是：下单、支付、发货、确认收货。但并不是所有的交易都会走此流程，比如买家下单之后不支付或关闭订单。针对这种非线性过程，处理情况主要有以下几种：(1）业务过程的统一

   我们以流程结束标志为依据，关闭订单也是结束标志，统一起来。

   (2）针对业务关键里程碑构建全面的流程

   对于没有支付或没有发货的交易订单也将其纳入流程来，相关的业务字段置孔。

   (3）循环流程的处理

   主要解决问题是一个业务过程有多个日期。使用业务过程的第一次发生日期还是最近发生日期，根据用户决定。

2. **多源过程** 针对多源业务建模，主要考虑事实表的粒度问题。

3. **业务过程取舍** 当拥有大量的业务过程时，模型的实现复杂度会增加，特别是对于多源业务过程，模型的精合度过高，此时需要根据商业用户需求，选取关键的里程碑。

物理实现

逻辑模型和物理模型密不可分，针对累积快照事实表模型设计，其有不同的实现方式。**第一种：增量存储** 以业务实体的结束时间分区。即每周期仅处理增量部分的数据，针对状态无变化的数据比较适合；**第二种：全量快照** 状态有变化，每天的分区存储昨天的全量数据和当天的增量数据合并的结果，对于数据量在可控范围内的情况可以采用如下 保存策略: 如果存储空间和成本可接受，完整存储，确保能够追溯到历史每天数据状态 存储空间有限，考虑移动历史快照数据到冷盘，需要使用的时候可恢复 数据历史状态数据无太大价值，可以考虑部分删除，比如近保留每月最后一天的快照数据；**第三种：拉链** 针对于全量表的变化形式，数据量大、但缓慢变化、需要跟踪历史状态，和缓慢渐变维类似。

设计准则

同事务事实表设计一样

表 11.7　三种事实表的比较

	事务事实表	周期快照事实表	累积快照事实表
时期/时间	离散事务时间点	以有规律的、可预测的间隔产生快照	用于时间跨度不确定的不断变化的工作流
日期维度	事务日期	快照日期	相关业务过程涉及的多个日期
粒度	每行代表实体的一个事务	每行代表某时间周期的一个实体	每行代表一个实体的生命周期
事实	事务事实	累积事实	相关业务过程事实和时间间隔事实
事实表加载	插入	插入	插入与更新
事实表更新	不更新	不更新	业务过程变更时更新

05

无事实的事实表

在维度模型中，事实表用事实来度量业务过程，不包含事实或度量的事实表称为**无事实的事实表。**虽然没有明确的事实，但可以用来支持业务过程的度量。常见的无事实的事实表主要有如下两种：

第一种是事件类的，记录事件的发生。

如阿里巴巴数据仓库中，最常见的是日志类事实表。

第二种是条件、范围或资格类的，记录维度与维度多对多之 间的关系。

如客户和销售人员的分配情况、产品的促销范围等。

06

聚集型事实表

数据仓库的性能是数据仓库建设是否成功的重要标准之一。聚集主要是通过汇总明细粒度数据来获得改进查询性能的效果。通过访问聚集数据，可以减少数据库在响应查询时必须执行的工作量，能够快速响应用户的查询，同时有利于减少不同用户访问明细数据带来的结果不一致问题。如阿里巴巴将使用频繁的公用数据，通过聚集进行沉淀，比如卖家最近 l 天的交易汇总表、卖家最近 N 天的交易汇总表、卖家自然年交易汇总表等。这类聚集汇总数据，被叫作**公共汇总层。**

相对于明细事实表，聚合事实表通常是在明细数据表的基础上，按照一定的粒度粗细进行的汇总、聚合操作，它的粒度较明细数据粒度粗，同时伴随着细节信息的丢失;在数仓层次结构中，通常位于dws层，一般作为通用汇总数据存在，也可以是更高粒度的指标数据。

基本原则

- **一致性** 聚集表必须提供与查询明细粒度数据一致的查询结果。
- **避免单一表设计** 不要在同一个表中存储不同层次的聚集数据；否则将会导致双重计算或出现更糟糕的事情。
- **聚集粒度可不同** 聚集并不需要保持与原始明细粒度数据一样的粒度，聚集只关心所需要查询的维度。

基本步骤

Step 1：确定聚集维度。

Step 2：确定一致性上钻。

Step 3：确定聚集事实。

常见聚集型事实表

数据仓库中，按照日期范围的不同，通常包括以下类别的聚集事实表

公共维度层-通用汇总

应对大部分可预期的、常规的数据需求，通常针对模式相对稳定的分析、BI指标计算、特征提取等场景，封装部分业务处理、计算逻辑，尽量避免用户直接使用底层明细数据，该层用到的数据范围比较广泛。

日粒度

主要应对模式稳定的分析、BI日报、特征提取场景，同时日粒度也为后续累积计算提供粗粒度的底层，数据范围一般为上一日的数据 。

周期性累积

主要应对明确的周期性分析、BI周期性报表，数据范围一般在某个周期内的

历史累积

顾名思义，历史以来某一特定数据的累积，通常在用户画像、经营分析、特征提取方面场景较多，设计数据范围比较广泛，通常是计算耗时较长的一部分，比如某门店累积营业额、某用户累积利润贡献、用户首次下单时间(非可度量、描述性)。

聚集补充说明

1. **聚集是不跨越事实的**

   聚集是针对原始星形模型进行的汇总，为了获取和查询与原始模型一致的结果，聚集的维度和度量必须与原始模型保持一致，因 此聚集是不跨越事实的。

2. **聚集带来的问题**

   聚集会带来查询性能的提升，但聚集也会增加 ETL 维护的难度。当子类目对应的一级类目发生变更时，先前存在的、已经被汇总到聚集表中的数据需要被重新调整。这一额外工作随着业务复杂性的增加，会导致多数 ETL 人员选择简单强力的方法，删除并重新聚集数据。