> 关注公众号：大数据技术派，回复资料，领取1024G资料。

雪花模型和星型模型

前面我们在学习数仓建模—建模方法论的时候提到了雪花模型和星型模型以及星座模型的这个概念，但是也对这个概念进行了一定的解释，但是那一片是为了介绍方法论，所以重点还是在方法论上面，这一节我们单独介绍一下这几个模型。

我们知道我们采用的是维度建模，模型的实现主要指的是在维度建模过程中，需要对维度表和事实表进行关联设计，而这里我们对维度表的设计，就决定了我们最终与事实表关联的之后的形态。也就是说我们可以根据事实表和维度表的关系，又可将常见的模型分为星型模型和雪花型模型以及星座模型

星型模型和雪花模型的**主要区别在于对维度表的拆分，对于雪花模型，维度表的设计更加规范，一般符合3NF**；而**星型模型，一般采用降维的操作，利用冗余来避免模型过于复杂，提高易用性和分析效率。**

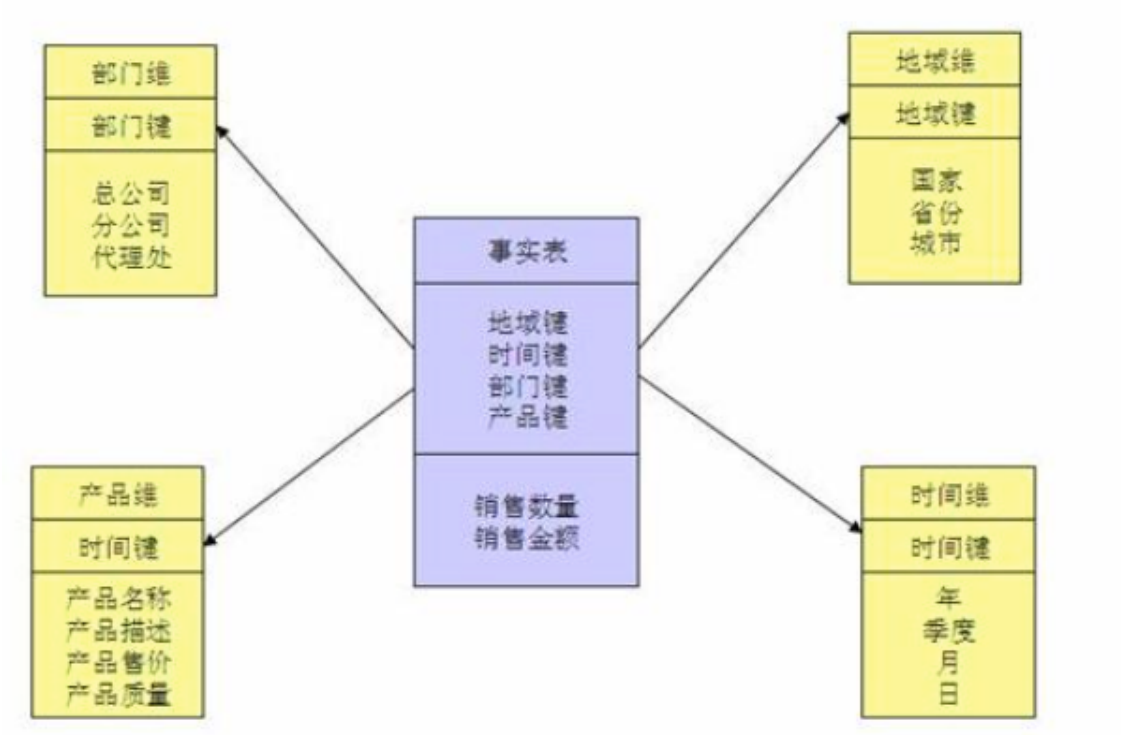
星型模型

核心是**一个事实表及多个非正规化描述的维度表**组成，维度表之间是没有关联的，维度表是直接关联到事实表上的，只有当维度表极大，存储空间是个问题时，才考虑雪花型维度，简而言之，最好就用星型维度即可。

维度表是直接关联到事实表上的，维度表之间是没有关联的这其实就说明了**维度表只有一层，中心只有一个那就是事实表**，当所有维表都直接连接到“ 事实表”上时，整个图解就像星星一样，故将该模型称为星型模型。

其实很多人都记不住，你可以这样理解，你可以这样想象，你从来都不会看到夜空里的星星两个是挨在一起的，都是孤零零的一个

星型架构是一种非正规化的结构，多维数据集的每一个维度都直接与事实表相连接，不存在渐变维度，所以数据有一定的冗余，如在地域维度表中，存在国家 A 省 B 的城市 C 以及国家 A 省 B 的城市 D 两条记录，那么国家 A 和省 B 的信息分别存储了两次，即存在冗余。



星型模是一种多维的数据关系，它由一个事实表和一组维表组成。每个维表都有一个维作为主键，所有这些维的主键组合成事实表的主键。强调的是对维度进行预处理，将多个维度集合到一个事实表，形成一个宽表。

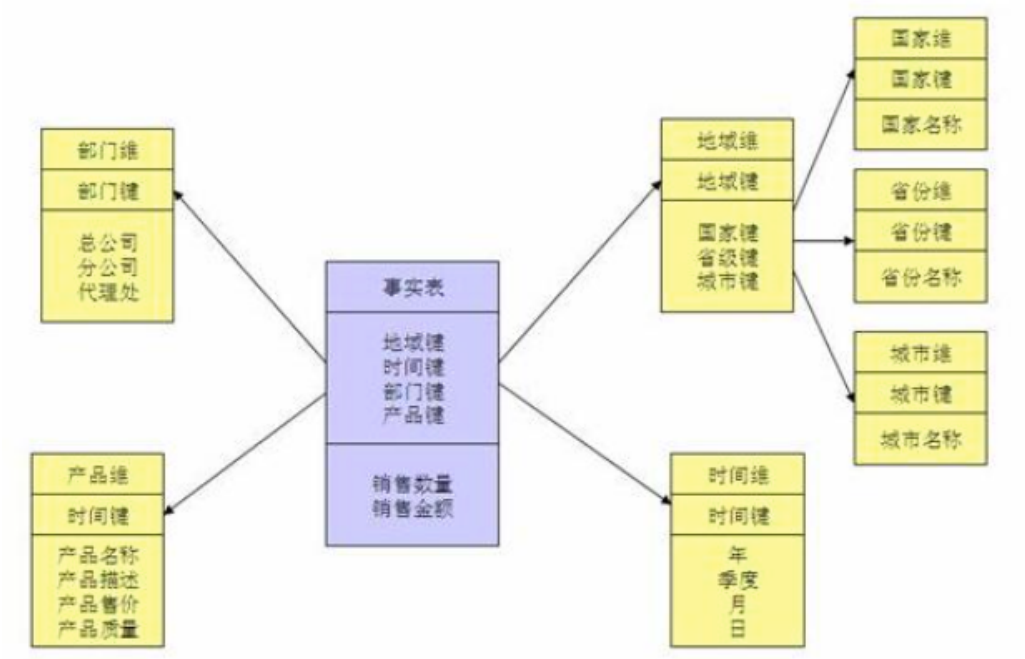
这也是我们在使用hive时，经常会看到一些大宽表的原因，大宽表一般都是事实表，包含了维度关联的主键和一些度量信息，而维度表则是事实表里面维度的具体信息，使用时候一般通过join来组合数据，相对来说对OLAP的分析比较方便。

雪花模型

星形模式中的维表相对雪花模式来说要大，而且不满足规范化设计。雪花模型相当于将星形模式的大维表拆分成小维表，满足了规范化设计。然而这种模式在实际应用中很少见，因为这样做会导致开发难度增大，而数据冗余问题在数据仓库里并不严重，但是需要注意的实业务系统一般就是这样设计的，因为这样的设计对于数据更新和模块解耦很方便。

当有一个或多个维表没有直接连接到事实表上，而是通过其他维表连接到事实表上时，其图解就像多个雪花连接在一起，故称雪花模型，可以认为雪花模型是星型模型的一个扩展，它对星型模型的维表进一步层次化，原有的各维表可能被扩展为小的事实表，形成一些局部的 "层次 " 区域，这些被分解的表都连接到主维度表而不是事实表。

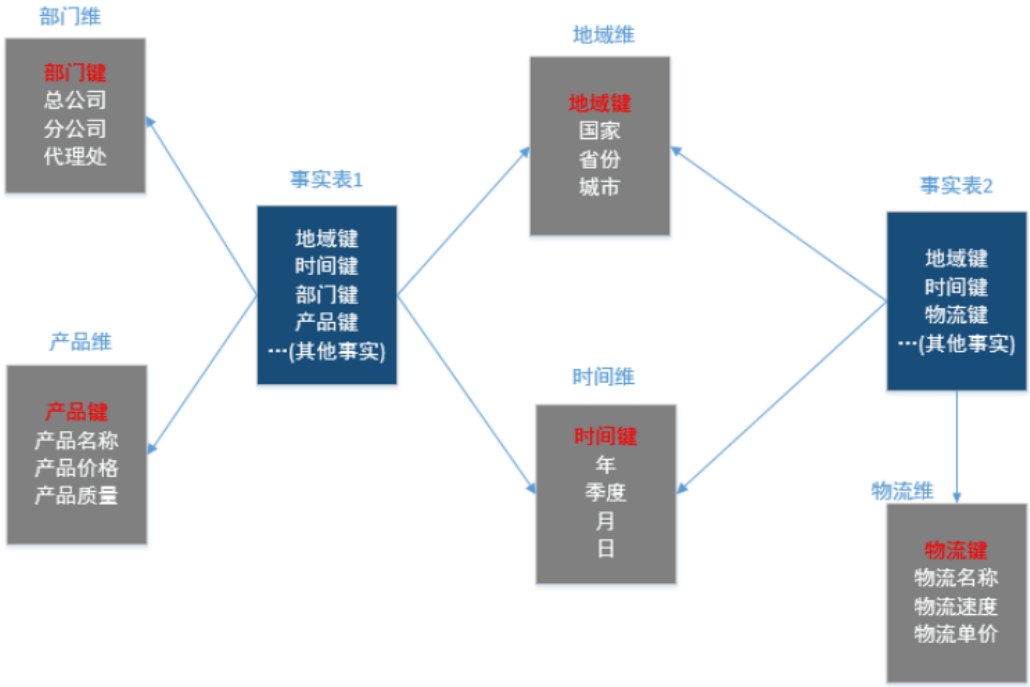
雪花模型更加符合数据库范式，减少数据冗余，但是在分析数据的时候，操作比较复杂，需要join的表比较多所以其性能并不一定比星型模型高。



星座模型

前面介绍的两种维度建模方法都是多维表对应单事实表，但在很多时候维度空间内的事实表不止一个，而一个维表也可能被多个事实表用到。在业务发展后期，绝大部分维度建模都采用的是星座模式。

可以认为是多个事实表的关联或者是星型模型的关联，其实到了业务发展后期都是星座模型



总结

我们看一下雪花模型和星型模型的对比

属性	星型模型	雪花模型
数据总量	多	少
可读性	容易	差
表个数	少	多
查询速度	快	慢
冗余度	高	低
对实时表的情况	增加宽度	字段比较少，冗余底
扩展性	差	好

星型模型的设计方式主要带来的好处是能够提升查询效率，因为生成的事实表已经经过预处理，主要的数据都在事实表里面，所以只要扫描实时表就能够进行大量的查询，而不必进行大量的join，其次维表数据一般比较少，在join可直接放入内存进行join以提升效率，除此之外，星型模型的事实表可读性比较好，不用关联多个表就能获取大部分核心信息，设计维护相对比较简答。

雪花模型的设计方式是比较符合数据库范式的理念，设计方式比较正规，数据冗余少，但在查询的时候可能需要join多张表从而导致查询效率下降，此外规范化操作在后期维护比较复杂。

通过上面的对比，我们可以发现数据仓库大多数时候是比较适合使用星型模型构建底层数据Hive表，通过大量的冗余来提升查询效率，星型模型对OLAP的分析引擎支持比较友好，这一点在Kylin中比较能体现。而雪花模型在关系型数据库中如MySQL，Oracle中非常常见，尤其像电商的数据库表。在数据仓库中雪花模型的应用场景比较少，但也不是没有，所以在具体设计的时候，可以考虑是不是能结合两者的优点参与设计，以此达到设计的最优化目的。

猜你喜欢

数仓建模—宽表的设计

Spark SQL知识点与实战

Hive计算最大连续登陆天数

Hadoop 数据迁移用法详解

Flink计算pv和uv的通用方法