数仓高内聚低耦合是怎么做的

定义

- 高内聚：强调模块内部的相对独立性，要求模块内部的元素尽可能的完成一个功能，不混杂其他功能，从而使模块保持简洁，易于理解和管理。

- 低耦合：模块之间的耦合度要尽可能的低，避免模块之间的复杂依赖，使得每个模块都可以独立存在，从而减少模块间的相互影响，提高系统的可维护性。

做到低耦合、高内聚

一般复杂的公共逻辑可以采用抽象类和抽象方法的方式下沉到共有模块中，然后由相关子类去实现抽象方法，来实现不同的功能。这样可以将复杂的逻辑拆分成各个子类，使得类之间的耦合度降低，提高代码的可维护性。

数仓中多重粒度的作用，以及实现

定义

在数据仓库中，粒度是指数据的细度。粒度越高，表示数据越细致，每个数据点所包含的信息量也就越大。粒度越低，表示数据的概括性越强，每个数据点所包含的信息量也就越小。

在数据仓库中，多重粒度指的是将数据按照多个不同的粒度进行存储，以便在需要时更方便地进行查询和分析。例如，可以将数据按年、月、日等不同的粒度进行存储，以便根据需里对数据进行按年、按月、按日等不同维度的分析。多重粒度数据仓库在实际应用中非常常见，能够满足大多数数据分析的需求。

作用

多重粒度数据仓库可以让我们更方便地对数据进行分析和查询，具体有以下几点作用:

1.提高查询效率: 将数据按照多个不同粒度存储，可以让我们更快地找到所需的数据。例如，如果我们需要查询某一天的销售数据，直接查询按日粒度存储的数据即可，而不用扫描整个数据仓库。

2、减少数据冗余:在数据仓库中，将数据按照多个粒度存储，可以减少数据几余，节省空间。例如，如果我们将每一天的销售数据都单独存储，那么一年的数据就需要存储 365 天的数据;如果将每一月的销售数据存储，则一年的数据只需要存储 12 个月的数据。

3、方便数据分析:多重粒度数据仓库可以让我们更方便地对数据进行分析。例如，如果我们想要对某一天的销售数据进行分析，可以直接查询按日粒度存储的数据;如果想要对某-月的销售数据进行分析，可以直接查询按月粒度存储的。

实现

在数据仓库中实现多重粒度是指在数据仓库中设计多种方式来表示和存储时间相关的数据。这样就可以在不同的粒度(例如年、月、日、小时等)》上查询数据，从而满足不同的分析需求

常用的实现方式有两种: 1.时间维度表:将时间的不同粒度分别建立为单独的维度表，并与事实表进行关联。例如，可以建立年、月、日、小时等维度表，并通过外键关联到事实表中。

2.时间层级表:将时间的不同粒度存储在同一个表中，并设计为层级结构。例如，可以将时间表设计为“年-月-日-小时”的层级结构，将每个时间点都存储在同一个表中。

具体选择哪种方式，取决于业务需求和数据查询的频率。

时间维度表的优势在于查询速度快，但维护成本较高，需要单独维护多个表。

时间层级表的优势在于维护成本低，但查询速度可能较慢。

如何提高查询效率

1. 优化数据库结构，统一管理所有数据，减少查询的次数;
2. 使用缓存技术，将查询结果保存到内存中，加速查询;
3. 合理利用索引，提高查询的效率;
4. 采用分布式系统，将查询任务分发到多台机器，提高查询速度;

5. 采用消息队列技术，将批量数据进行拆分，减少查询时间；

6. 利用数据库定时备份技术，减少查询时间；

7. 采用数据库分片技术，将数据分布到多个数据库，提高查询效率；

8. 采用数据库视图技术，将复杂的SQL语句拆分为多个简单的SQL语句，提高查询效率；

9. 采用SQL优化技术，充分利用数据库的索引，提高查询效率；

10. 采用数据库集群技术，将数据分布到多个数据库服务器，提高查询效率；

数仓数据域划分几种方式

我们采用四种方式对数仓数据域进行划分：

1. 按照业务类型划分：比如销售、财务、研发、物流等等。

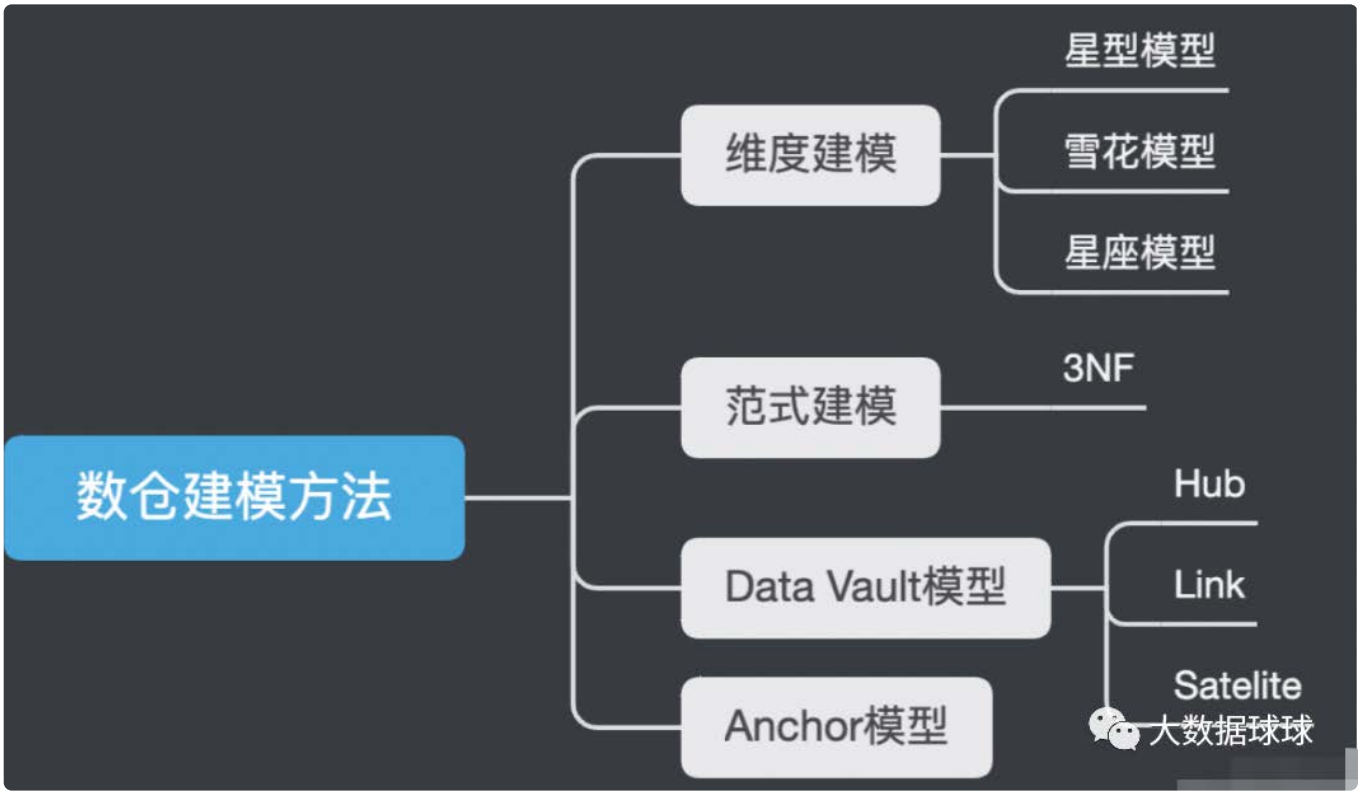
2. 根据需求方划分:比如需求方为财务部，就可以设定对应的财务主题域，而财务主题域里面可能就会有员工工资分析，投资回报比分析等主题。

3 按照功能或应用划分:比如微信中的朋友圈数据域、群聊数据域等，而朋友圈数据域可能就会有用户动态信息主题、广告主题等。

4 按照部门划分:比如可能会有运营域、技术域等，运营域中可能会有工资支出分析、活动宣传效果分析等主题.

数仓构建有几种方式

建模方式



数仓构建方式

> 说法1：

1. 集成数仓构建：这种方法把各类数据存储在各自的数据库中，然后通过数据集成工具将数据集成到一个数仓中，以满足数据分析的需求。

2. 数据融合数仓构建：这种方法是在源数据层面进行融合，将源数据经过整合、清洗、转换等操作，构建数据仓库，以满足数据分析的需求。

3. 元数据数仓构建：这种方法是在元数据层面进行数据整合，建立元数据的抽象层，以便更好的管理和操作数据，以满足数据分析的需求。

4. 大数据数仓构建：这种方法是将大数据仓库采用分布式存储的方式进行数据存储，以满足数据分析的需求。

> 说法2

1、基于现有系统构建：利用现有系统，如ERP、SCM、CRM等，通过开发定制或者引入第三方软件，构建数字化仓库管理系统。

2、新建系统构建：从零开始，根据实际需求，开发建立一套新的数字化仓库管理系统。

3、集成构建：将现有的传统仓库管理系统和新的数字化仓库管理系统进行整合，构建全新的仓库管理系统。

> 说法3

数仓构建有多种方式。这取决于您的需求、技术基础、数据来源和构建目标。

常用的数仓构建方式包括:

1.基于 ETL的数仓构建:在这种方式中，您可以使用 ETL (提取、转换、加载)工具来从源系统提取数据，然后在数仓中进行转换和加载。

2.基于 ELT 的数仓构建:在这种方式中，您可以使用 ELT(提取、加载、转换)工具来将数据从源系统提取到数仓，然后在数仓中进行转换。

3、基于事件驱动的数仓构建: 在这种方式中，您可以使用事件驱动的架构，在事件发生时即时地将数据加载到数仓中

4.基于流的数仓构建:在这种方式中，您可以使用流处理框架，以流的方式将数据实时加载到数仓中。

哪种方式最合适，取决于您的业务需求和技术环境。

粒度操作

常用的粒度操作有上卷、下钻、切片、切块、旋转、拉伸、锯齿等。

- 上卷：上卷指的是增加粒度，将原来比较细的粒度提升到更大的粒度，从而让整体更清晰，更容易理解，更容易把握。

- 下钻：下钻指的是减小粒度，将原来比较粗的粒度放低到更细的粒度，从而更加细致的把握数据的细节，更加清楚的把握数据的特征。
- 切片：切片指的是将数据分割成若干个数据片，从而更加方便地进行管理和操作。
- 切块：切块指的是将数据分割成若干个数据块，从而更加方便地进行管理和操作。
- 旋转：旋转指的是对数据进行旋转操作，从而让数据看起来更加美观，更容易理解。
- 拉伸：拉伸指的是对数据进行拉伸操作，从而让数据看起来更加清晰，更容易理解。
- 锯齿：锯齿指的是对数据进行锯齿操作，从而让数据看起来更加精细，更容易理解。

SQL实现

SQL实现上卷下钻切片切块旋转通常包括以下步骤：

1、首先使用SELECT语句从数据库中选择所需要的表；

2、使用GROUP BY语句将数据按照指定的维度进行分组；

3、使用HAVING语句按照指定条件对分组数据进行筛选；

4、使用ORDER BY语句对分组后的数据进行排序；

5、使用LIMIT和OFFSET语句对数据进行分页；

6、使用CASE语句对数据进行旋转以形成报表。

数仓中ODS层命中多少为合理

数仓中ODS层的命中率是指数据在ODS层中的命中率。一般来说，ODS层的命中率应该尽量高，因为这意味着更多的数据是从ODS层获取的，而不是从原始数据源获取，这样可以减少对原始数据源的访问压力，并提高数据获取的效率。

不过，ODS层的命中率也不能太高，因为如果ODS层的命中率过高，就意味着ODS层的数据不够新，这可能会导致ODS层的数据不够准确。所以，ODS层的命中率应该适中，一般来说，在80%~90%左右是合理的。

当然，ODS层的命中率也受到很多因素的影响，比如ODS层的数据更新频率、ODS层的数据容量等。因此，具体的合理命中率还需要根据实际情况具体分析。

数仓价值链的体现和实现。

数仓价值链的体现主要是通过以下几个方面：

1、数据采集：数仓系统要能够从各种来源采集数据，包括传统数据库、网络日志、企业应用系统和第三方数据源等。

2、数据清洗：数据采集后，可能存在脏数据、缺失数据等情况，数仓系统要对数据进行清洗，使其符合分析的要求。

3、数据存储：将清洗后的数据存储到数仓系统中，以便后续的分析和查询。

4、数据分析：使用数仓系统中的数据进行分析，提供对决策者有价值的信息。

5、数据报告：将分析结果呈现给决策者，帮助他们做出决策。

建立数仓通常需要经过以下步骤

建立数据仓库通常要经过以下几个步骤:

1、需求分析：在建立数据仓库之前，需要先进行需求分析，确定数据仓库的目的和功能，并规划数据仓库的架构和设计。

2、数据清洗和整合：在建立数据仓库之前，需要对来源数据进行清洗和整合，以确保数据的准确性和完整性。

3、构建数据模型：根据数据仓库的需求和功能，构建数据仓库的逻辑数据模型。

4、建立物理数据模型：根据逻辑数据模型，建立物理数据模型，并根据需要设计数据仓库的存储结构。

5、数据加载：将来源数据加载到数据仓库中。

6、数据分析和报告：使用数据仓库中的数据进行分析和生成报告，为企业决策提供依据。

7、维护和优化：对数据仓库进行定期的维护和优化，以确保数据的准确性和完整性。

指标生命周期可以从哪几个方面来评估

指标从被创建到被废弃的整个过程。指标生命周期可以从以下几个方面来评估:

1. 创建时间: 指标被创建的时间点
2. .更新频率:指标数据更新的频率，包括实时更新、每日更新、每周更新等
3. .使用频率:指标被使用的频率，包括每日使用、每周使用、每月使用等。
4. 使用场景:指标被使用的场景，包括决策支持、规划、监控等。
5. .相关性: 指标与业务的相关性，即指标能否反映业务状态
6. .准确性: 指标数据的准确性，即指标能否反映实际情况
7. 可解释性: 指标数据的可解释性，即指标能否被正确理解和解释
8. 可操作性: 指标能否被有效地操作，即指标数据能否被用于实际的决策或行动。

通过对指标生命周期的评估，可以帮助企业更好地管理和使用指标，提高指标的有效性和价值

数据治理在做什么

数据治理是一种指导和管理数据生命周期的框架和方法。这包括数据的收集、存储、处理、使用和保护。

数据治理的目的是提高数据质量，并确保数据在组织内被合理使用。数据治理可以帮助组织有效地使用数据，并防止数据泄露或滥用。

做数仓的目的

数据仓库（Data Warehouse）是一种存储大量历史数据的系统，它主要用于数据分析和报告。数据仓库通常包含来自多个不同来源的数据，并使用ETL（提取，转换和加载）过程将数据转换为可以进行分析的形式。

数据仓库的目的是为管理层提供一个在线的数据分析工具，使他们能够快速获取有关公司业务的信息，并基于这些信息做出决策。数据仓库的建立是为了满足企业决策的需要，为企业的经营决策、规划决策、计划决策和控制决策提供依据，即为企业决策供给。

数据仓库是数据集成的基础，也是数据挖掘的前提。因此，建立数据仓库的目的不仅仅是为了供给决策，还包括为数据挖掘和数据分析提供基础。

以上仅供参考，有什么问题可以联系我一起学习。