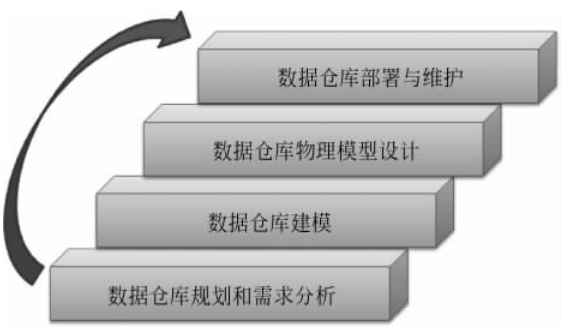
第 3 章　数据仓库设计

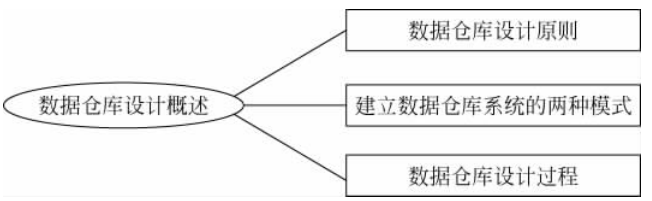


数据仓库设计步骤

本章指南

3.1　数据仓库设计概述

知识梳理



3.1.1　数据仓库设计原则

数据仓库设计是建立一个面向企业决策者的分析环境或系统。数据仓库的设计原则是以业务和需求为中心，以数据来驱动。前者是指围绕业务方向性需求、业务问题等，确定系统范围和总体框架；后者是指其所有数据均建立在已有数据源基础上，从已存在于操作型环境中的数据出发进行数据仓库设计。

3.1.2　建立数据仓库系统的两种模式

1. 先整体再局部的构建模式

该构建模式最早由 W. H. Inmon 提出，先创建企业数据仓库，即对分散于各个业务数据库中的数据特征进行分析，在此基础上实施数据仓库的总体规划和设计、构建一个完整的数据仓库、提供全局数据视图，再从数据仓库中分离部门业务的数据集市，即逐步建立针对各主题的数据集市，以满足具体的决策需求。

这种构建模式通常在技术成熟、业务过程理解透彻的情况下使用，也称为自顶向下模式，如图 3.1 所示，其中数据由数据仓库流向数据集市。

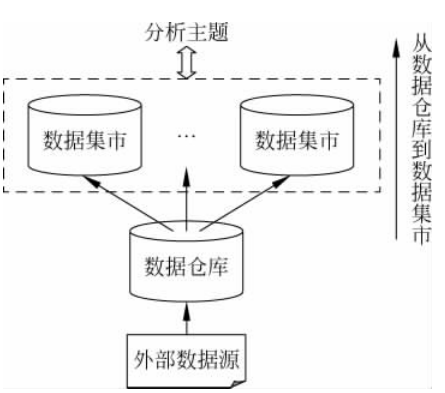


图 3.1　自顶向下模式

该模式的优点是数据规范化程度高，由于面向全企业构建了结构稳定和数据质量可靠的数据中心，可以相对快速有效地分离面向部门的应用，从而最小化数据冗余与不一致性；当前数据、历史数据与详细数据整合，便于全局数据的分析和挖掘。

其缺点是建设周期长、见效慢，风险程度相对大。

2. 先局部再整体的构建模式

该构建模式最早由 Ralph Kimball 提出，是先将企业内各部门的要求视作分解后的决策子目标，并针对这些子目标建立各自的数据集市，在此基础上对系统不断进行扩充，逐步形成完善的数据仓库，以实现对企业级决策的支持。

这种构建模式也称为自底向上模式，如图 3.2 所示，其中数据由数据仓库流向数据集市。

该模式的优点是投资少、见效快；在设计上相对灵活；由于部门级数据的结构简单，决策需求明确，因此易于实现。

其缺点是数据需要逐步清洗，信息需要进一步提炼，如数据在抽取时有一定的重复工作，还会有一定级别的冗余和不一致性。

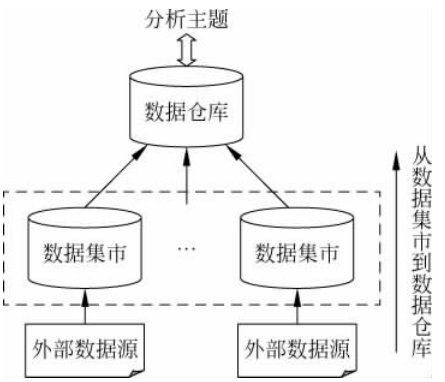


图 3.2 自底向上模式

3.1.3 数据仓库设计过程

数据仓库的设计从数据、技术和应用三方面展开，各方面工作完成之后，进行数据仓库部署，然后数据仓库投入运行使用，同时管理人员对数据仓库进行维护，完成数据仓库的一个生命周期，如图 3.3 所示。

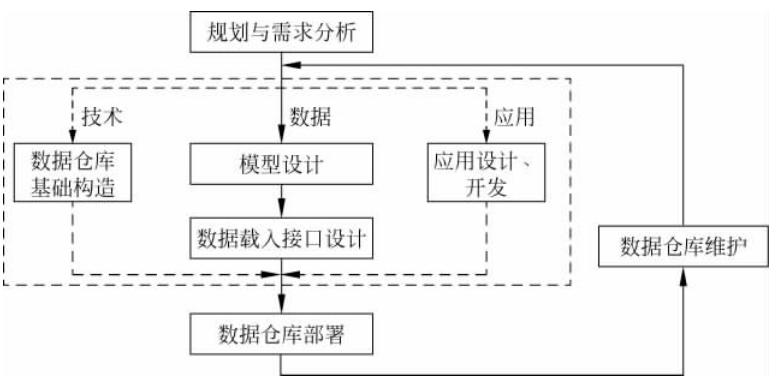


图 3.3 数据仓库建立的基本框架

技术路线的实施分为技术选择和产品选择两个步骤。如何采用有效的技术和合适的开发工具是实现一个好的数据仓库系统的基本条件。

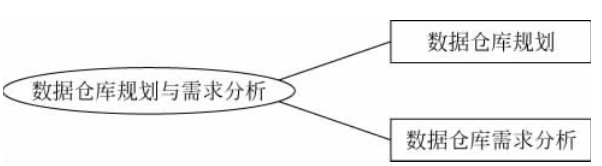
数据路线的实施可以分为模型设计、物理设计和数据处理三个步骤，用以满足对数据的有效组织和管理。

应用路线的实施分为应用设计和应用开发两个步骤。数据仓库的建立最终是为应用服务的，所以需要对应用进行设计和开发，以更好地满足用户的需要。

其中数据路线的实施是后面讨论的重点。

3.2 数据仓库规划与需求分析

知识梳理



3.2.1　数据仓库规划

数据仓库的规划主要产生建设数据仓库的策略规划，确定建立数据仓库的长期计划，并为每一建设阶段设定目标、范围和验证标准。数据仓库的策略规划包括：

(1) 明确用户的战略远景、业务目标。

(2) 确定建设数据仓库的目的和目标。

(3) 定义清楚数据仓库的范围、优先顺序、主题和针对的业务。

(4) 定义衡量数据仓库成功的要素。

(5) 定义精简的体系结构、使用技术、配置、容量要求等。

(6) 定义操作数据和外部数据源。

(7) 确定建设所需要的工具。

(8) 概要性地定义数据获取和质量控制的策略。

(9) 数据仓库管理及安全。

其中非常重要的一条就是业务目标，建设数据仓库的目的就是通过集成不同的系统信息为企业提供统一的决策分析平台，帮助企业解决实际的业务问题(例如，如何提高客户满意度和忠诚度，降低成本、提高利润，合理分配资源，有效进行全面绩效管理等)。因此在规划数据仓库时要以应用驱动，充分考虑如何满足业务目标。

数据仓库体系结构的建设将是一个系统工程。它的规划、设计、开发、投产、改造将是一个循环往复、长时期的工作，数据仓库的建设过程中应该遵循：在大中心的模式下，实现信息集中管理、统筹规划、整体设计、分步实施的原则。同时，在系统实施过程中要体现“统一规划、统一标准、统一选型、统一开发”的“四统一原则”。建成的数据仓库体系结构应满足以下几点。

(1) 全面的：必须满足企业各管理职能部门的业务需求，提供全套产品，提供服务与支持，以及拥有能提供补充产品的合作伙伴。所有这些，才能确保数据仓库能满足现在及将来的特殊要求。一个全面的解决方案是在技术基础上的延伸，包括分析应用，从而使业务人员能真正从数据仓库系统中获益，提高企业运作效率、扩大市场以及平衡两者间的关系。

(2) 完整的：必须适合现存的环境，它必须提供一个符合工业标准的完整的技术框架，以保证系统的各个部分能协调一致地工作。

(3) 不受限制的：必须适应变化，必须能迅速、简单地处理更多的数据及服务更多的用户，以满足不断增长的需求。

(4) 最优的：必须在企业受益、技术及低风险方面经过验证，必须在市场上保持领先地位，具有明显的竞争优势和拥有大量的合作伙伴产品。

3.2.2　数据仓库需求分析

数据仓库的特点是面向主题，按主题组织数据。所谓主题就是分析决策的目标和要求，因此主题是建立数据仓库的前提。数据仓库应用系统的需求分析，必须紧紧围绕着主题来进行，主要包括主题分析、数据分析和环境要求分析。

1. 主题分析

需求分析的中心工作是主题分析，主题是由用户提出的分析决策的目标和需求，它有宏观和微观等多种形式。在此阶段要通过开发方与用户方大量的沟通，把用户提出的需求进行梳理，归纳出主题并分解成若干需求层次，构成从宏观到微观、从综合到细化的主题层次结构。

对于每个主题，需要进行详细的调研，确定要分析的指标和用户从哪些角度来分析数据即维度（包括维度层次），还要确定用户分析数据的细化或综合程度即粒度。

主题、指标、维度和粒度是建立数据仓库的基本要素。

2. 数据分析

数据仓库系统以数据为核心，因此数据的分析非常重要。在确定了分析主题后，就需要从业务系统的数据源入手，进行数据的分析。数据分析包括以下的工作。

（1）数据源分析：分析目前存在哪些数据源，这些数据源能否支撑主题的需要，了解清楚这些数据源的结构、数据之间的关系，并给出详细的描述。

（2）数据数量分析：数据仓库对数据数量有一定的最低要求，对数据密度、宽度都有一定的要求，因此需要分析数据源的数据能否达到这些要求。

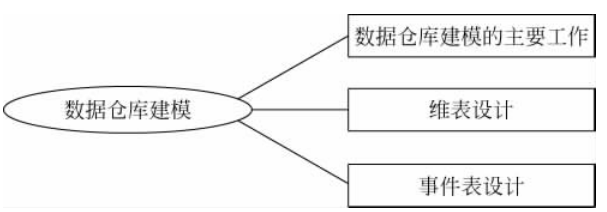
（3）数据质量分析：需要对数据源的数据质量进行分析，确定数据的正确性、一致性、规范性和全面性能否达到要求。

3. 环境要求分析

需要对满足需求的系统平台与环境提出要求，包括设备、网络、数据、接口、软件等的要求。

3.3 数据仓库建模

知识梳理



3.3.1 数据仓库建模的主要工作

数据仓库建模是指设计数据仓库的逻辑模型。逻辑建模是数据仓库实施中的重要一环，因为它能直接反映出业务部门的需求，同时对系统的物理实施有重要的指导作用。

数据仓库的建模主要是确定数据仓库中应该包含的数据类型及其相互关系，其主要工作如下。

1. 确定主题域

主题是在较高层次上将企业信息系统中的数据进行综合、归类和分析利用的一个抽象概念，每一个主题基本对应一个宏观的分析领域。在逻辑意义上，它是对应企业中某一宏观分析领域所涉及的分析对象。

主题域是对某个主题进行分析后确定其边界。确定主题边界实际上需要进一步理解业务关系，因此在设计好主题后，还需要对这些主题进行初步的细化才便于获取每一个主题应该具有的边界，如图 3.4 所示是确定主题域的示意图。在设计数据仓库时，一般是一次先建立一个

主题或企业全部主题中的一部分，因此在大多数数据仓库的设计过程中都有一个主题域的选择过程。主题域的确定必须由最终用户和数据仓库的设计人员共同完成。

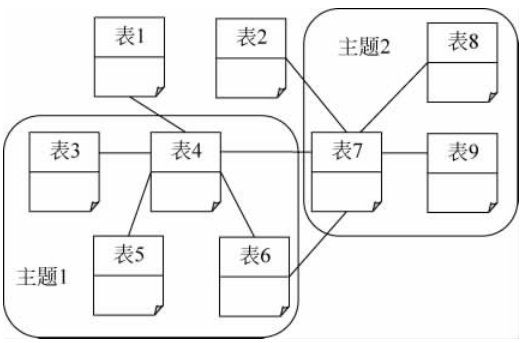


图 3.4　确定主题域的示意图

在确定系统所包含的主题域后，对每个主题域的内容进行较详细的描述，描述的内容包括主题域的公共码键、主题域之间的联系和代表主题的属性组。

例如，对于一个电子商务数据仓库，管理层需要分析的主题一般包括商品主题、客户主题和销售主题，如表 3.1 所示是其中商品、销售和顾客主题的详细描述。

表 3.1　主题的详细描述

主题名	公共键	属　性　组
商品	商品编号	商品基本信息：商品编号，商品名称，类型等
销售	订单编号	销售基本信息：订单编号，日期，顾客编号，商品编号，销售数量、销售金额等
		销售评价信息：订单编号，顾客编号，商品编号，评语，打分等
顾客	顾客编号	顾客基本信息：顾客编号，姓名，性别，年龄，学历，住址，电话等

2. 粒度设计

粒度问题是设计数据仓库的一个最重要的方面。粒度是指数据仓库的数据单位中保存数据的细化或综合程度的级别。细化程度越高，粒度级就越小；相反，细化程度越低，粒度级就越大。如图 3.5 所示是数据粒度的示意性表示。

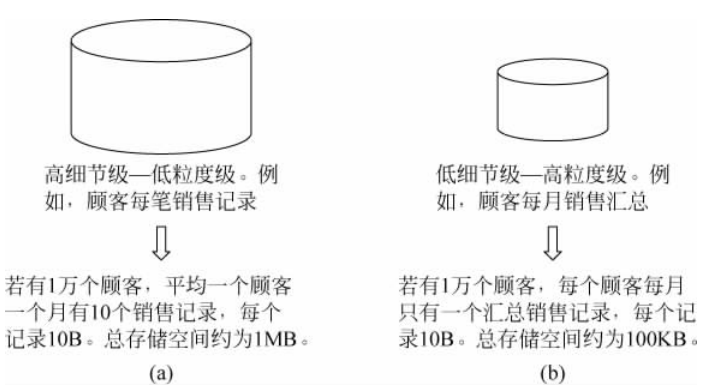


图 3.5　数据的粒度

数据的粒度一直是一个设计问题。在操作型系统中，几乎总是选择最低粒度级。但在数据仓库环境中，对粒度并没有统一的规定，需要设计者根据实际情况来确定数据的粒度级别。

在数据仓库环境中粒度之所以是主要的设计问题，是因为它深深地影响存放在数据仓库中的数据量的大小，同时影响数据仓库所能回答的查询类型。

例如，如果粒度太大，数据量可能比较少，但得不到更详细的查询结果。当数据仓库中仅仅存放顾客每月的销售汇总数据时，就不能按日期和星期分析顾客的购物情况了。所以，粒度设计就是在数据仓库中的数据量大小与查询的详细程度之间做出权衡。

3. 数据仓库建模

通常采用维度建模法为数据仓库建模，主要内容是确定数据仓库的多维数据模型是星形模型、雪花模型还是事实星座模型。在此基础上设计相应的维表和事实表，从而得到数据仓库的逻辑模型。

例如，对于第 2 章图 2.11 的数据仓库概念模型，采用关系数据库时，其逻辑模型描述如下（下划线部分是关系的主键）：

时间维表（Time_id，日期，年份，季度，月份，周）

地点维表（Locate_id，街道，城市，省，国家）

商品维表（Item_id，商品名，品牌，分类）

顾客维表（Customer_id，顾客名，顾客住址，顾客类型）

销售事实表（Time_id，Item_id，Locate_id，Customer_id，销售量，销售金额）

从使用的效率角度考虑，设计数据仓库时要考虑以下因素：

（1）尽可能使用星形架构，如果采用雪花结构，还需要进一步规范化维表。

（2）维表的设计应该符合通常意义上的范式约束，维表中不要出现无关的数据。

（3）事实表中包含的数据应该具有必需的粒度。

（4）对事实表和维表中的关键字必须创建索引。

（5）保证数据的引用完整性，避免事实表中的某些数据行在聚集运算时没有参加进来。

有关维表和事实表的详细设计将在后面章节中介绍。

4. 确定数据分割策略

分割是指把逻辑上是统一整体的数据分割成较小的、可以独立管理的物理单元进行存储，以便能分别处理，从而提高数据处理的效率。

数据分割为什么如此重要呢？因为在管理数据时小的物理单元比大的物理单元具有更大的灵活性，包括更容易重构、索引、顺序扫描、重组、恢复和监控等。如果是大块的数据，就达不到访问数据的灵活性要求。因而，对所有当前细节的数据仓库数据都要进行分割。

分割可以按时间、地区、业务类型等多种标准来进行，也可以按自定义标准，如图 3.6 所示采用的是按时间分割数据。但在多数情况下，数据分割采用的标准不是单一的，而是多个标准的组合。

选择适当的数据分割标准，一般要考虑以下几方面的因素：

（1）数据量大小。

（2）数据分析处理的实际情况。

（3）简单易行。

（4）与粒度的划分策略相统一。

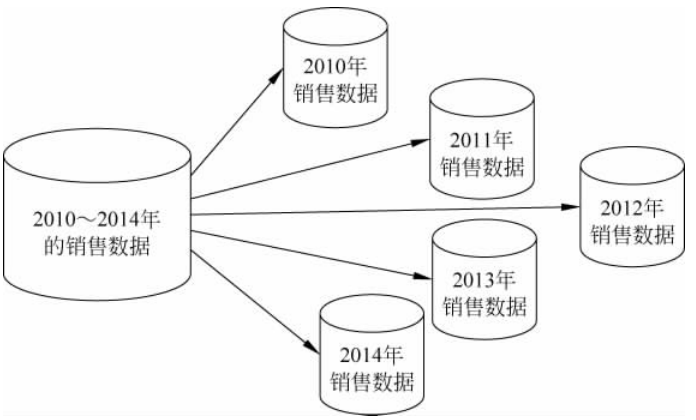


图 3.6　按时间分割数据

(5) 数据的稳定性。

3.3.2　维表设计

1. 维表的特征

维表用于存放维信息,包括维属性(列)和维成员。一个维用一个维表表示。维表通常具有以下数据特征:

(1) 维通常使用解析过的时间、名字或地址元素,这样可以使查询更灵活。例如时间可分为年份、季度、月份和时期等,地址可用地理区域来区分,如国家、省、市、县等。

(2) 维表通常不使用业务数据库的关键字作为主键,而是对每个维表另外增加一个额外的字段作为主键来识别维表中的对象。在维表中新设定的键也称为代理键。

(3) 维表中可以包含随时间变化的字段,当数据集市或数据仓库的数据随时间变化而有额外增加或改变时,维表的数据行应有标识此变化的字段。

2. 维的类型

维表中维的类型包括结构维、信息维、分区维、分类维、退化维、一致维和父子维多种类型。

(1) 结构维。结构维表示在维层次结构组成中的信息度量,如年份、月份和日期可以组成一个结构维,商品销售地点可以组成另一个结构维,由此可以分析某个时期在某个地区销售价的商品总量。

(2) 信息维。信息维是由计算字段建立的。用户也许想通过销售利润了解所有产品的销售总额。也许希望通过增加销售来获得丰厚的利润。然而,如果某一款商品降价销售,可能会发现销售量虽然很大,而利润却很小或几乎没有利润。从另一方面看,用户可能希望通过提高某种产品的价格获得较大的利润。这种产品可能具有较高的利润空间,但销量却可能很低。因此,就利润建立一个维,包括每种商品利润和全部利润的维,就销售总量建立一个度量,这样可以提供有用的信息,这个维就是一个信息维。

(3) 分区维。分区维是以同一结构生成的两个或多个维。例如,对于时间维,每一年有相同的季度、相同的月和相同的天(除了闰年以外,而它不影响维),在 OLAP 分析中,将频繁使用时间分区维来分割数据仓库中的数据,其中一个时间维中的数据是针对 2014 年的,而另一个时间维中的数据是针对 2015 年的,建立事实表时,可以把度量分割为 2014 年的数据和

2015年的数据，这会提高分析性能。

(4) 分类维。分类维是通过对一个维的属性值分组而创建的。如果顾客表中有家庭收入属性，那么，可能希望查看顾客根据收入的购物方式。为此，可以生成一个含有家庭收入的分类维。例如，如果有以下家庭每年收入的数据分组：0～20000元、20001～40000元、40001～60000元、60001～100000元和大于100001元。

(5) 一致维。当有好几个数据集市要合并成一个企业级的数据仓库时，可以使用一致维来集成数据集市以便确定所有的数据集市可以使用每个数据集市的事实。所以，一致维常用于属于企业级的综合性数据仓库，使得数据可以跨越不同的模式来查询。

(6) 父子维。父子维度基于两个维度表列，这两列一起定义了维度成员中的沿袭关系。一列称为成员键，标识每个成员；另一列称为父键，标识每个成员的父代。该信息用于创建父子链接，该链接将在创建后组合到代表单个元数据级别的单个成员层次结构中。父子维度用通俗的话来讲，就是这个表是自反的，即外键本身就是引用的主键。例如，公司组织结构中，分公司是总公司的一部分，部门是分公司的一部分，员工是部门的一部分，通常公司的组织架构并非处在等层次上，例如总公司下面的部门看起来就和分公司是一样的层次。因此父子维的层次通常是不固定的。

在数据仓库的逻辑模型设计中，有一些维表是经常使用的，它们的设计形成了一定的设计原则，如时间维、地理维、机构维和客户维等，所以在设计维表时应遵循这些设计原则。又例如，数据仓库存储的是系统的历史数据，业务分析最基本的维度就是时间维，所以每个主题通常都有一个时间维。

3. 维表中的概念分层

维表中的维一般包含层次关系，也称为概念分层，如在时间维上，按照“年份－季度－月份”形成了一个层次，其中年份、季度、月份成为这个层次的三个级别。

概念分层的作用如下：

(1) 概念分层为不同级别上的数据汇总，如上卷操作提供了一个良好的基础。

(2) 综合概念分层和多维数据模型的潜力，如下钻操作可以对数据获得更深入的洞察力。

(3) 通过在多维数据模型中，在不同的维上定义概念分层，使得用户在不同的维上从不同的层次对数据进行观察成为可能。

(4) 多维数据模型使得从不同的角度对数据进行观察成为可能，而概念分层则提供了从不同层次对数据进行观察的能力；结合这两者的特征，可以在多维数据模型上定义各种OLAP操作，为用户从不同角度不同层次观察数据提供了灵活性。

3.3.3 事实表设计

1. 事实表的特征

事实表是多维模型的核心，是用来记录业务事实并做相应指标统计的表，同维表相比，事实表具有如下特征：

(1) 记录数量很多，因此事实表应当尽量减小一条记录的长度，避免事实表过大而难于管理。

(2) 事实表中除度量外，其他字段都是维表或中间表(对于雪花模式)的关键字(外键)。

(3) 如果事实相关的维很多，则事实表的字段个数也会比较多。

2. 事实表的类型

事实表的粒度能够表达数据的详细程度。从用途的不同来说，事实表可以分为以下三类。

(1) 原子事实表：是保存最细粒度数据的事实表，也是数据仓库中保存原子信息的场所。

(2) 聚集事实表：是原子事实表上的汇总数据，也称为汇总事实表。即新建立一个事实表，它的维度表是比原维度表要少，或者某些维度表是原维度表的子集，如用月份维度表代替日期维度表；事实数据是相应事实的汇总，即求和或求平均值等。

(3) 合并事实表：是指将位于不同事实表中处于相同粒度的事实进行组合建模而成的一种事实表。即新建立一个事实表，它的维度是两个或多个事实表的相同维度的集合，事实是几个事实表中感兴趣的事实。合并事实表的粒度可以是原子粒度也可以是聚集粒度。

聚集事实表和合并事实表的主要差别是合并事实表一般是从多个事实表合并而来的。但是它们的差别不是绝对的，一个事实表既是聚集事实表又是合并事实表是很有可能的。因为一般合并事实表需要按相同的维度合并，所以很可能在做合并的同时需要进行聚集，即粒度变粗。注意维度和事实表应在同一个粒度上。

3. 聚集函数

在查询事实表时，通常使用到聚集函数，一个聚集函数从多个事实表记录中计算出一个结果。如一个事实表中销售量是一个度量，如果要统计所有的销售量，便用求和聚集函数，即SUM(销售量)。在设计事实表时需要为每个度量指定相应的聚集函数。度量可以根据其所用的聚集函数分为以下三类。

(1) 分布的聚集函数：将这类函数用于 n 个聚集值得到的结果和将函数用于所有数据得到的结果一样。例如 COUNT(求记录个数)、SUM(求和)、MIN(求最小值)、MAX(求最大值)等。

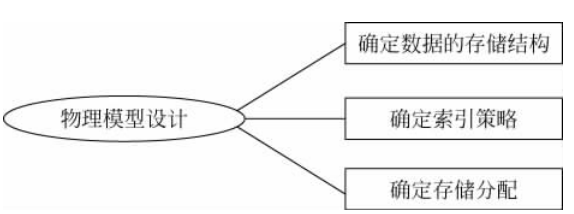
(2) 代数的聚集函数：函数可以由一个带 m 个参数的代数函数计算(m 为有界整数)，而每个参数值都可以由一个分布的聚集函数求得，例如 AVG(求平均值)等。

(3) 整体的聚集函数：描述函数的子聚集所需的存储没有一个常数界，即不存在一个具有 m 个参数的代数函数进行这一计算，例如 MODE(求最常出现的项)。

在设计事实表时，可以利用减少字段个数、降低每个字段的大小和把历史数据归档到单独事实表中等方法来减小事实表的大小。

3.4　数据仓库物理模型设计

知识梳理



数据仓库的物理模型是逻辑模型在数据仓库中的实现模式。构建数据仓库的物理模型与所选择的数据仓库开发工具密切相关。这个阶段所做的工作是确定数据的存储结构，确定索引策略和确定存储分配等。

设计数据仓库的物理模型时，要求设计人员必须做到以下几方面：

(1) 要全面了解所选用的数据仓库开发工具，特别是存储结构和存取方法。

(2) 了解数据环境、数据的使用频度、使用方式、数据规模以及响应时间要求等，这些是对时间和空间效率进行平衡和优化的重要依据。

(3) 了解外部存储设备的特性，如分块原则、块大小的规定、设备的 I/O 特性等。

3.4.1 确定数据的存储结构

一个数据仓库开发工具往往都提供多种存储结构供设计人员选用，不同的存储结构有不同的实现方式，各有各的适用范围和优缺点。设计人员在选择合适的存储结构时应该权衡三个方面的主要因素：存取时间、存储空间利用率和维护代价。

同一个主题的数据并不要求存放在相同的介质上。在物理设计时，常常要按数据的重要程度、使用频率以及对响应时间的要求进行分类，并将不同类的数据分别存储在不同的存储设备中。重要程度高、经常存取并对响应时间要求高的数据就存放在高速存储设备上，如硬盘；存取频率低或对存取响应时间要求低的数据则可以放在低速存储设备上，如磁盘或磁带。此外，还可考虑如下策略。

1. 合并表组织

在常见的一些分析处理操作中，可能需要执行多表连接操作。为了节省 I/O 开销，可以把这些表中的记录混合放在一起，以减少表连接运算的代价，这称为合并表组织。这种组织方式在访问序列经常出现或者表之间具有很强的访问相关性时具有很好的效果。

2. 引入冗余

在面向某个主题的分析过程中，通常需要访问不同表中的多个属性，而每个属性又可能参与多个不同主题的分析过程。因此可以通过修改关系模式把某些属性复制到多个不同的主题表中，从而减少一次分析过程需要访问的表的数量。

3. 分割表组织

在逻辑设计中按时间、地区、业务类型等多种标准把一个大表分割成许多较小的、可以独立管理的小表，称为分割表。这些分割表可以采用分布式的存储方式，当需要访问大表中的某类数据时，只需访问分割后的对应小表，从而提高访问效率。

4. 生成导出数据

在原始、细节数据的基础上进行一些统计和计算，生成导出数据，并保存在数据仓库中，避免在分析过程中执行过多的统计和计算操作，提高分析的性能，又避免不同用户进行重复统计可能产生的偏差。

3.4.2 确定索引策略

数据仓库的数据量很大，因而需要对数据的存取路径进行仔细的设计和选择。由于数据仓库的数据都是不常更新的，因而可以设计多种多样的索引结构来提高数据存取效率。

在数据仓库中，设计人员可以考虑对各个数据存储建立专用的、复杂的索引，以获得最高

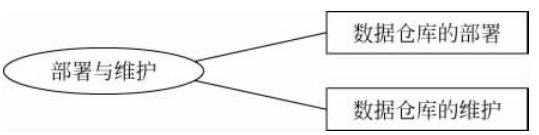
的存取效率，因为在数据仓库中的数据是不常更新的，也就是说每个数据存储是稳定的，因而虽然建立专用的、复杂的索引有一定的代价，但一旦建立就几乎不需维护索引的代价。

3.4.3　确定存储分配

许多数据仓库开发工具提供了一些存储分配的参数供设计者进行物理优化处理，例如，块的尺寸、缓冲区的大小和个数等，它们都要在物理设计时确定。这同创建数据库系统时的考虑是一样的。

3.5　数据仓库部署与维护

知识梳理



3.5.1　数据仓库的部署

完成前面各项工作之后，可以进入数据仓库的部署阶段，主要包括用户认可、初始装载、桌面准备和初始培训。

1. 用户认可

用户的认可在部署阶段不只是一个形式而是绝对必需的，在关键用户没有对数据仓库表示满意前不要强行进行部署。用户是否认可主要通过相关测试来进行，下面是测试的一些要点：

(1) 在每个主题域或部门，让用户选择几个典型的查询和报表，执行查询并产生报表，最后从操作型系统生成报表作为验证数据库产生的报表。

(2) 测试预定义查询和报表。

(3) 测试 OLAP 系统。让用户选择大约 5 个典型分析会话进行测试并与操作型系统的结果比较。

(4) 进行前端工具的可用性设计测试。

(5) 如果数据仓库支持 Web，则需要进行 Web 特性测试。

(6) 进行系统性能测试。

2. 初始装载

初始装载的主要任务是运行接口程序，将数据装入到数据仓库中。初始装载的主要步骤如下：

(1) 删除数据仓库关系表中的索引。因为初始装载数据量很大，建立索引耗费大量的时间。

(2) 可以限制关系完整性的检验。

(3) 确保已经建立合适的检查点。为了避免在装载过程中失败需要全部重新开始装载，

所以必须建立检查点。

(4) 装载维表。

(5) 装载事实表。

(6) 基于已经为聚集和统计表建立的计划,建立基于维表和事实表的聚集表。

(7) 如果装载时停止了索引建立,那么现在建立索引。

(8) 检查数据装载参考完整性约束。在装载过程中,所有的参考性错误记录在系统中,检查日志文件,找出所有装载异常。

3. 桌面准备

桌面准备的主要工作是安装好所有需要的桌面用户工具,包括桌面计算机需要的硬件、网络连接的全部需求,测试每个客户的计算机。

4. 初始培训

培训用户学习数据仓库相关的概念、相关的内容和数据访问工具,建立对初始用户的基本使用支持。

3.5.2 数据仓库的维护

维护数据仓库的工作主要是管理日常数据装入的工作,包括刷新数据仓库的当前详细数据、将过时的数据转化成历史数据、清除不再使用的数据、管理元数据等。

另外,还有如何利用接口定期从操作型环境向数据仓库追加数据、确定数据仓库的数据刷新频率等。

练 习 题

1. 单项选择题

(1) 有关数据仓库的开发特点,下列说法(　　)是不正确的。

A. 数据仓库开发要从数据出发

B. 数据仓库使用的需求在开发出来后才会明确

C. 数据仓库开发是一个不断循环的过程

D. 数据仓库中数据的分析和处理十分灵活,没有固定的开发模式

(2) 关于数据仓库设计,下列说法中正确的是(　　)。

A. 不可能从用户的需求出发来进行数据仓库的设计

B. 只能从各部门业务应用的方式来设计数据模型

C. 在进行数据仓库主题数据模型设计时要强调数据的集成性

D. 在进行数据仓库概念模型设计时,必须要设计实体关系图

(3) 有关数据仓库粒度设计的叙述中正确的是(　　)。

A. 粒度越细越好　　B. 粒度越粗越好

C. 粒度应该与数据仓库的主题相对应　　D. 以上都不对

(4) 有关数据仓库分割策略的叙述中正确的是(　　)。

A. 分割越细越好

B. 分割策略与数据量大小和速度等因素有关

C. 分割越粗越好

D. 以上都不对

(5) 有关数据仓库建模的叙述中正确的是(　　)。

A. 因为需求分析中已经考虑主题,建模时不再需要确定主题域

B. 因为需求分析中已经确定项目的所有功能,没有必要再进行数据仓库建模工作

C. 数据仓库建模是设计概念模型,继而导出逻辑模型

D. 数据仓库建模是设计物理模型

(6) 有关数据仓库物理模型设计的叙述中正确的是(　　)。

A. 存储结构中不能存在任何数据冗余

B. 尽可能多地建立索引

C. 尽可能把在逻辑上关联的数据放在一个表中

D. 以上都不对

2. 问答题

(1) 简述数据仓库设计的步骤。

(2) 简述维有哪些类型。

(3) 简述事实表有哪些类型。

(4) 简述数据仓库物理模型设计的主要内容。