CHAPTER 3

第 3 章　数据仓库系统的设计与开发

通过前两章的介绍，我们对数据仓库的概念、体系结构与存储结构、ETL过程等内容都有了一定的了解，那么如何建立一个数据仓库系统？在实际工作中数据仓库系统是如何设计和开发出来的？特别是在数据仓库创建以后，又怎样在此基础上建立多维数据模型？这些问题实际上属于数据仓库的实践范畴，与具体的应用系统环境特别是所使用的数据库环境、开发工具甚至开发人员都有密切的关系。本章以微软公司的 SQL Server 2005 为应用开发环境，通过实例介绍数据仓库的设计与开发过程。

3.1　数据仓库系统的设计与开发概述

3.1.1　建立数据仓库系统的步骤

数据仓库系统的建立在一定程度上说，是一个复杂甚至漫长的过程，因为数据仓库系统的开发涉及到源数据系统、数据仓库对应的数据库系统及数据分析与报表工具等诸多应用问题。因此，数据仓库系统的创建不是一蹴而就的，将数据从原有的操作型业务环境移植到数据仓库环境本身就是一项复杂而艰巨的工作。一般来说，一个数据仓库系统的建立需要经过如下步骤。

（1）收集和分析业务需求。用户需求往往不确定，在数据仓库环境中，决策支持分析人员往往是企业或事业组织的中上层管理人员，他们对决策分析的需求不能预先做出规范说明。他们对开发人员说："让我看看能得到什么，然后我才能告诉你我真正需要什么"。因此，数据仓库应该在海量的数据中为用户提供有用、及时、全面的信息，以帮助用户做出正确的决策。

（2）建立数据模型和数据仓库的物理设计。通过设计数据仓库的概念模型、逻辑模型和物理模型，可以得到企业或事业数据的完整而清晰的描述信

息。数据仓库的数据模型通常是面向主题建立的，同时又为多个面向应用的
数据源的集成提供了统一的标准。数据仓库的核心内容包括组织的各个主题域、主题域之间的联系、描述主题的码和属性组等。

(3) 定义数据源。也叫做定义记录系统(system of records)，往往会形成一个操作型数据存储区，数据仓库中的数据来源于多个已有的操作型业务系统。一方面，各个系统的数据都是面向应用的，不能完整地描述企业中的主题域；另一方面，多个数据源的数据之间存在着许多不一致，如命名、结构和单位不一致等，甚至数据的内容也可能不一致。所以必须在已有系统中定义记录系统。记录系统是一个内容正确、在多个数据源间起决定作用的操作型数据源。它的特点是：数据最完整、最准确、最及时，结构最适合于数据仓库，并且与外部数据源最为接近。

(4) 选择数据仓库技术和平台。技术和平台选型对建设数据仓库来说非常重要，而且一旦选定，在数据仓库系统实施完成后将很难改变，平台及技术的切换成本非常高，所以选型一定要充分重视和高度谨慎。

(5) 从操作型数据库中抽取、清洗及转换数据到数据仓库。本部分内容参见第 2 章。

(6) 选择访问和报表工具，选择数据库连接软件，选择数据分析和数据展示软件。根据用户的具体情况及其分析需求和数据量大小等因素选择。

(7) 更新数据仓库。确定数据仓库的更新策略，开发或配置数据仓库更新子系统，实现数据仓库数据的自动更新。

3.1.2　数据仓库系统的生命周期

数据仓库系统的开发与设计是一个动态的反馈和循环过程。一方面，数据仓库的数据内容、结构、粒度、分割以及其他物理设计根据用户所返回的信息需要不断地调整和完善，以提高系统的效率和性能。另一方面，通过不断地理解需求，使得最终用户能做出更准确、更有用的决策分析。

一个数据仓库系统包括两个主要部分：一是数据仓库数据库，用于存储数据仓库的数据；二是数据分析应用系统，用于对数据仓库数据库中的数据进行分析。因此，数据仓库系统的设计也包括数据仓库数据库的设计和数据仓库应用的设计两个方面。事实上，系统的设计开发是基于数据仓库的规划、需求分析及数据模型建立等前期工作的，数据仓库系统在经过分析与设计两个重要阶段后，就会进入数据仓库系统的实施阶段，实施完成后便转入系统维护阶段。在系统的使用和维护过程中，用户会提出新的需求，同时也会有新技术出现，因此数据仓库系统在用户使用评价和新需求确认的基础上，进入新一轮的分析、设计开发、实施与维护的循环。这个开发与使用过程是一个不断循环、完善和提高的过程。在一般情况下，一个数据仓库系统不可能在一个循环过程中完成，而是经过多次循环开发，每次循环都会为系统增加新的功能，使数据仓库的应用得到新的提高。图 3.1 示意了这个循环的开发过程，这个过程也叫数据仓库系统的生命周期。

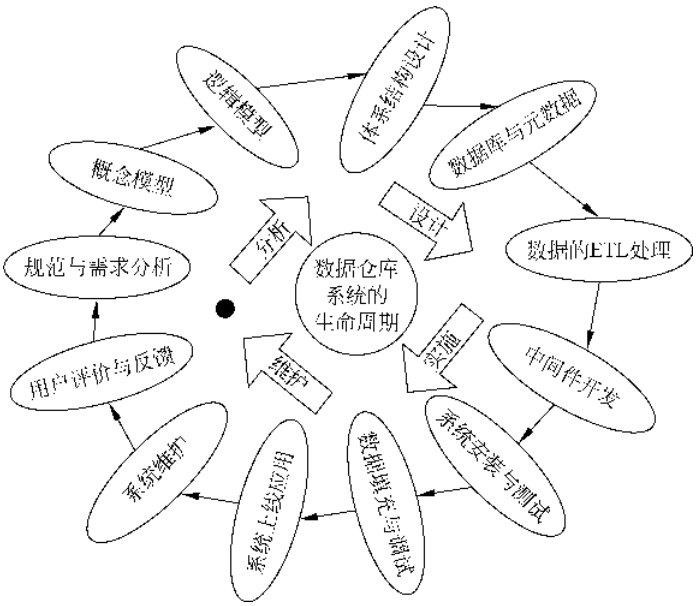

图 3.1　数据仓库系统的生命周期

3.1.3　建立数据仓库系统的思维模式

1. 自顶向下(top-down)

这种模式首先把 OLTP 数据通过 ETL 汇集到数据仓库中,然后再把数据通过复制的方式推进各个数据集市中,其优点如下。

(1) 数据来源固定,可以确保数据的完整性。

(2) 数据格式与单位一致,可以确保跨越不同数据集市进行分析的正确性。

(3) 数据集市可以保证有共享的字段。因为都是从数据仓库中分离出来的。

2. 自底向上(bottom-up)

这种模式首先将 OLTP 数据通过 ETL 汇集到数据集市中,然后通过复制的方式提升到数据仓库中,其优点如下。

(1) 首先构建数据集市的工作相对简单,易成功。

(2) 这种模式也是实现快速数据传送的原型。

3.1.4　数据仓库数据库的设计步骤

数据仓库数据库的设计如图 3.2 所示,主要工作包括收集、分析和确认业务分析需求,分析和理解主题和元数据、事实及其量度、粒度和维度的选择与设计、数据仓库的物理存储方式的设计等。

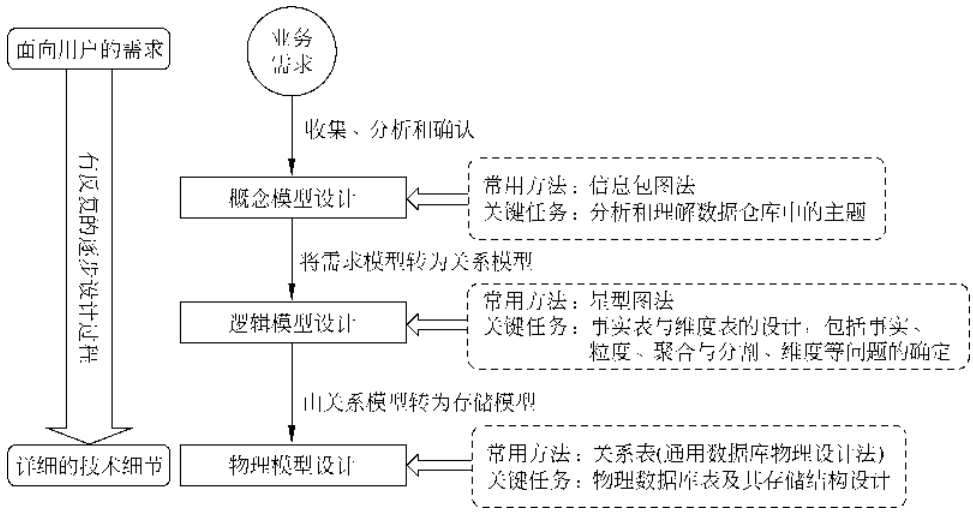

图 3.2　数据仓库数据库设计示意图

3.2　基于 SQL Server 2005 的数据仓库数据库设计

SQL Server 2005 集成了三个服务来实现数据仓库系统的开发：SQL Server 2005 Analysis Services、SQL Server 2005 Integration Services 和 SQL Server 2005 Reporting Services，同时还提供了一个数据仓库与商业智能应用系统的开发环境——SQL Server Business Intelligence Development Studio。它们的关系如图 3.3 所示。

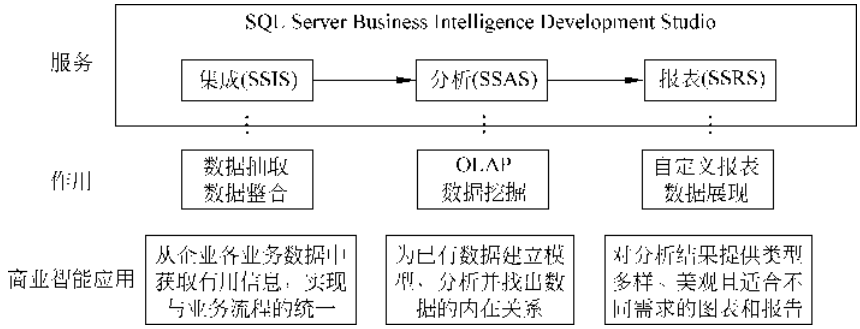

图 3.3　MS SQL Server 2005 的数据仓库架构

(1) SSAS(分析服务)提供了所有业务数据的统一整合视图，可以作为传统报表、在线分析处理、关键性能指示器(key performance indicators，KPI)记分卡和数据挖掘的基础。同时提供了一个元数据模型以满足不同需求。其中的所有多维数据集和维度定义都可从统一空间模型(UDM)中查阅。UDM 是一个中心元数据库，其中定义了业务实体、业务逻辑、计算和度量，可作为所有报表、电子表格、OLAP 浏览器、KPI 和分析应用程序的源来使用。在 SQL Server 2005 中，关系数据库和多维数据库之间的界限变得模糊。可以将数据存储在关系数据库或多维数据库中，还可以使用主动缓存功能，充分利用两种数据库各自的优点。

(2) SSIS(集成服务)具有出色的 ETL 和整合能力,提供了构建企业级 ETL 应用程序所需的功能和性能,使得组织机构能更加容易地管理来自于不同的关系型和非关系型数据源的数据。SSIS 是可编程的、可嵌入的和可扩展的,这些特性使其成为理想的 ETL 平台。

(3) SSRS(报表服务)是一个基于服务器的完整报告平台,可创建、管理和交付传统报告和交互式报告。它包括创建、分发和管理报告所需的一切工具和信息。同时,其标准模块化设计和应用程序编程接口(API)使软件开发人员、数据提供商和企业能够集成原有系统或第三方应用程序中的报表功能。

在 SQL Server 2005 中可以安装示例数据库 Adventure Works DW(数据仓库示例数据库),其主要数据都源于另一示例数据库 Adventure Works(OLTP 示例数据库)。

本节以 SQL Server 2005 作为数据仓库环境,来讲解数据仓库数据库的设计过程。示例数据是 SQL Server 2005 自带的示例数据库 Adventure Works 和 Adventure Works DW。

3.2.1 分析组织的业务状况及数据源结构

下面以 SQL Server 2005 示例数据库 Adventure Works DW 中所描述的 Adventure Works Cycles 公司的用户需求为例,介绍系统需求收集与分析过程。有关详细的业务信息参见 SQL Server 2005 有关数据仓库示例的帮助文档。

1. 公司概况

Adventure Works Cycles 是一家虚构的大型跨国制造公司,经营自行车及其相关配套产品,主要生产金属复合材料的自行车,产品远销北美、欧洲和亚洲市场。Adventure Works Cycles 公司总部设在华盛顿州的伯瑟尔市,雇佣了 500 名工人。此外,在 Adventure Works Cycles 市场中还活跃着一些地区销售团队。

2000 年,Adventure Works Cycles 购买了位于墨西哥的小型生产厂 Importadores Neptuno。Importadores Neptuno 为 Adventure Works Cycles 产品系列生产多种关键子组件,这些子组件将被运送到伯瑟尔市进行最后的产品装配。2001 年,Importadores Neptuno 转型成为专注于旅游登山车系列产品的制造商和销售商。

实现一个成功的会计年度之后,Adventure Works Cycles 现在希望通过以下方法扩大市场份额:专注于向高端客户提供产品、通过外部网站扩展其产品的销售渠道、通过降低生产成本来削减其销售成本。

2. 原材料采购、生产和销售等环节的业务流程介绍

(1) 原材料采购与仓储业务流程。

在公司内部由采购部负责原材料采购,采购部门下设一个经理和多个采购员。一种原材料有多个供应商,一个供应商可以提供多种原材料。原材料和供应商之间是多对多的关系。每个采购员负责多种原材料的采购,一种原材料只能由一个采购员来采购。采购员和商品之间是一对多的关系。采购员只需了解原材料和供应商的联系,而采购部门经理需要管理员工,并且还需要了解原材料的库存情况,以确定需要采购的商品并将任务分配给每个采购人员。

公司为了防止产品过分依赖于原材料价格,还需要对原材料进行批量存储,因此设立仓

库管理部门，专门负责原材料的存储管理，仓库管理部门管理多个仓库，下设一个经理和多个仓库管理员，每个仓库中拥有多个仓库管理员，每个管理员只能在一个仓库中进行工作。仓库管理员需要知道他所管理的仓库中存储的原材料的种类、数量、存储的时间、原材料的保值期及原材料进入仓库和离开仓库的时间等信息。一种原材料可以保存在多个仓库中，一个仓库可以保存多种原材料。仓库管理部门经理不但需要处理仓库管理员需要的数据，而且需要知道仓库管理员的基本信息（例如仓库管理员的家庭住址和电话等）。

（2）产品销售业务流程。

Adventure Works Cycles 的自行车及其相关产品远销北美、欧洲和亚洲市场。公司有网络销售和批发商销售两种销售渠道，因此，客户也分为两类，一类是从在线商店购买产品的消费者，通常是个人；一类是商店，即从 Adventure Works Cycles 销售代表处购买产品后进行转售的零售店或批发店。

对于销售部门，销售员关心的是商品的信息，即每种商品的价格、质量、颜色和规格等，以便向顾客推销相关的产品。因此，销售员最需要的数据就是商品的相关信息。销售部门经理一方面需要了解商品的销售情况，以便在某种商品缺货的时候通知仓储部门运送商品；另一方面，销售部经理还需要了解每个销售员的工作业绩，对每个销售员进行考核。因此销售部门经理需要了解商品、顾客和部门员工的情况。

3. 对数据源结构的分析与理解

从上面的分析可以看出，业务数据确实是多维的。不同部门对数据的需求不同，同一部门人员对数据需求也存在差异。如果考虑数据需求的层次问题，管理人员和不同的业务人员对数据要求的程度也各不相同。管理人员可能需要综合度较高或较为概括的数据，而业务人员需要细节数据。

实际上，对业务的理解是所有信息系统建设过程所需要的，只不过在设计数据仓库时需要从业务蕴涵的数据视角来理解业务。数据仓库是以历史数据为基础的，这一步本质上是理解这些历史数据的来源。

通常，操作型业务数据是数据仓库数据库的来源和基础，只有对它的内容足够了解和理解，才能很好地设计数据仓库和对数据进行 ETL 处理。

首先是要了解数据源（操作型业务数据）的结构，例如 Adventure Works 示例数据库把 Adventure Works Cycles 公司的业务数据分成 5 大部分，分别是表示人力资源的 Human Resources、表示人员信息如客户或供应商联系人等的 Person、表示产品信息的 Production、表示采购信息的 Purchasing 和表示销售信息的 Sales。

其次是要明确数据的内容，数据内容包括某个业务领域的数据表结构及其主外键关系，还包括各个数据表的具体字段构成情况。Adventure Works 示例数据库的业务数据内容简述如下。

（1）个人客户相关数据。个人客户是公司客户类型的一大类，即从 Adventure Works Cycles 在线商店购买产品的消费者。若 Sales. Customer 表的 CustomerType 列值为 I，则表示客户类型为个人，若为 S 则为批发商。与个人客户相关的表有 5 个，分别是：Person. Contact 表示客户的联系方式；Sales. Customer 表示客户的类型；Sales. Individual 表示个人客户的具体信息，其中 Demographics 列还以 XML 格式对个人客户的收入、爱好和车辆数目等进

行了统计；而对于客户的订单信息则放在 Sales. Sales Order Header 和 Sales. Sales Order Detail 两个表中。

（2）产品相关数据。Adventure Works Cycles 公司提供 4 类产品，包括公司自己生产的自行车、自行车零部件（替换件，如车轮、踏板或刹车等部件），还有从供应商处购买来转售给客户的自行车装饰和自行车附件等。和产品相关的表比较多，结构也较为复杂，产品相关的数据内容如表 3.1 所示。

表 3.1　产品（production）相关的表及其数据内容

数 据 表	数据表内容解释
Bill Of Materials	制造自行车和自行车子部件的所有零部件列表（BOM 结构），Product Assembly ID 列表示父级产品（即主产品），ComponentID 表示组装用的零件
Culture	列出使用了哪些语言来本地化产品说明
Location	列出产品和零件的库存位置
Product	由公司销售或用来制造自行车和自行车组件的各种产品信息
Product Category	产品分类，例如自行车或附件
Product CostHistory	列出不同时间点的产品成本
Product Description	列出各种语言的详细产品说明
Product Inventory	按地点统计的产品库存量
Product List Price History	列出不同时间点的产品价格
Product Model	与产品关联的产品型号
Product Model Product Description Culture	给出产品型号、产品说明及其本地化后的语言之间的交叉引用
Product Photo	列出所售产品的图像
Product Review	给出客户对产品的评价
Product Subcategory	产品类别的子类别

（3）原材料采购相关数据。采购部门购买自行车生产中需使用的原材料和零件，同时也购买一些产品直接转售，例如自行车装饰件和自行车附件，像水瓶和打气筒等。和原材料采购相关的表的数据内容如表 3.2 所示。

表 3.2　原材料采购（purchasing）相关的表及其数据内容

数 据 表	数据表内容
Person. Address	客户的通信地址信息
Person. Contact	供应商雇员的姓名，与 Vendor Contact 表相关联，将联系人映射到供应商。XML 数据类型的列 Additional ContactInfo 包含了联系人的其他联系方式（手机和传真等）
Product Vendor	将供应商与其提供的产品建立对应关系
Purchase Order Detail	采购订单的明细信息，包括订购的产品、数量和单价等
Purchase OrderHeader	采购订单的头信息，包括应付款总计、订购日期和订单状态等。Purchase OrderHeader 表与 Purchase Order Detail 表构成主—从关系
Ship Method	用于维护产品标准发货方法的查找表。Ship Method ID 列包含在 Purchase Order Header 表中

续表

数 据 表	数据表内容
Vendor	供应商的详细信息，例如供应商的名称和账号
Vendor Address	将客户链接到 Address 表中的地址信息。按类型对地址进行分类，例如开票地址、家庭住址和发货地址等。Address Type ID 列映射到 Address Type 表中
Vendor Contact	这是一个关联表。连接 Contact 和 Vendor 两个表

以上对比较重要的业务数据表进行了解释，其目的是提供一个可供项目参考的示例，若要完全理解业务数据，还要对操作型业务数据库中的表结构及表间关系进行深入的理解。这里以原材料采购数据中的表 Purchasing. Purchase Order Header 为例子分析此表的具体结构，如表 3.3 所示。从后续的 ETL 处理及业务分析和业务规则挖掘等操作中会发现，对这些业务数据表的理解和分析是相当重要的。因此，在实际项目实施中，应该对重要的业务数据表进行类似的分析。如表 3.3 所示。

表 3.3　Purchasing. Purchase Order Header 的表结构

列	数据类型	说　　明
Purchase Order ID	int	主键
Revision Number	tinyint	用于跟踪一段时间内采购订单变化的递增编号
Status	tinyint	订单状态：1＝等待批准，2＝已批准，3＝已拒绝，4＝完成
Employee ID	int	创建采购订单的雇员。指向 Employee. Employee ID 的外键
Vendor ID	int	采购订单所采购的产品的供应商。Vendor. Vendor ID 的外键
Ship Method ID	int	发货方法，Ship Method. Ship Method ID 的外键
Order Date	datetime	采购订单的创建日期
Ship Date	datetime	预计供应商的发货日期
Sub Total	money	采购订单小计
Tax Amt	money	税额
Freight	money	运费
Tota lDue	money	应付款(SubTotal＋TaxAmt＋Freight)
Modified Date	datetime	上次更新日期和时间

3.2.2　组织需求调研，收集分析需求

数据仓库应用系统不同于事务处理业务系统，其数据分析需求刚开始时并不十分明确，而数据仓库的数据来源往往来自各操作型业务数据库的历史数据和当期数据，因此，项目需求的收集与分析需要从历史数据与用户需求两个方面同时着手，采用“数据驱动＋用户驱动”的设计理念。

数据驱动是根据当前业务数据的基础和质量情况，以数据源的分析为出发点构建数据仓库。另一方面，用户驱动则是根据用户业务的方向性需求，从业务需要解决的具体问题出发，确定系统范围和需求框架，也叫需求驱动。

数据仓库的用户一般是企业或事业单位的管理者，在设计数据仓库系统时充分考虑用户的分析需求是十分必要的。同时，由于数据仓库的构建必须基于业务数据库等数据源，数

据源的结构也是不得不考虑的问题。如图 3.4 所示，常常采用“两头挤法”找出数据仓库系统的真正需求。

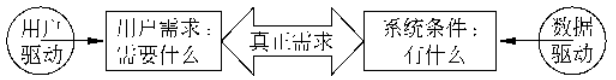

图 3.4　用户驱动与数据驱动相结合示意图

在 3.2.1 节的分析和理解业务数据库(Adventure Works)的过程中明确了企业的具体业务映射到数据库系统的细节，对从现有数据源中如何获取企业数据仓库需求对应的数据已经心中有数。现在我们从企业的各个视角对此企业数据仓库的分析需求作进一步的分析，发现企业需要且可以构造的主题。

实际上，企业每个部门都有观察企业业务的不同视角，这是需求多样性的一个方面。例如，对于 Adventure Works Cycles 公司来说，销售部门、采购部门和仓库管理部门等都有相应的视角，尽管这些业务是相关的，但是对数据的需求，特别是对分析数据的需求必然有所不同。

创建企业级数据仓库是一个面向企业各部门特别是管理部门的工程，但它的需求在前期可能是相当模糊的，因此，我们应该高度重视需求的调研与分析，尽量多地挖掘出用户的当前需求和潜在需求。

1. 关于用户需求的调研

用户方面，从组织机构的上层开始交流是非常有益的。上层行政官员可以提供许多令人惊奇的有关业务操作及其希望从该组织得到的内容。除此之外，还应该包括负责数据仓库项目或有关业务领域的行政职员，以及来自相关业务领域的负责向高级行政官员汇报的主管经理和为高级行政人员和主管经理准备报告的业务分析员。

根据不同的交谈对象，所提问题也应有所不同。很明显，针对高级行政人员和基层职员应该有不同的问题。通常也有些共性的问题域，例如：

(1) 他们的工作成绩是怎样得来的？即什么因素决定工作的成功与失败？

(2) 工作所需信息的分析过程需耗费多长时间？从这些分析中可做出什么决策？

(3) 信息分发的方式是什么？是报告、论文还是电子邮件？

(4) 怎样弥补信息的空缺？

(5) 分析这些数据需要哪一级的详细程度？

(6) 业务报表的来源是什么？谁对报告的制定、维护和分发负责？

对需求按照业务视角进行分类以后，可以进一步细化需求的提问，一般从业务目标、当前信息源、涉及的主题范围、关键性能指标和信息频率等方面分别提问。

(1) 业务目标：部门职责和目标是什么？怎样将这些目标融进企业的目标之中？要达到这些目标有哪些需要？成功的关键因素是什么？集团公司实现这些目标的障碍有哪些？需要购买外部数据吗？从哪里购买？

(2) 当前信息源和日常报表需求：在现有的日常报表过程中，当前传递了哪些信息？从何处获取这些分析数据？现在是如何加工处理的？这些信息的详细程度怎样？是太详细了，还是太粗略了？哪些操作会产生关于重要主题领域的数据和信息？

（3）主题领域：有关这方面的问题可以帮助确定对业务活动来说什么主题是很重要的。随着用户地位的不同，在他们的数据领域中，各种不同的领域或维度被明确地提出，这就使得对信息的整合变得容易了。这些问题集中在数据仓库中的数据应怎样被检索及用户怎样分析和筛选这些数据等。关于主题领域的问题如：

① 哪些维度或领域对数据的分析是有价值的？这些维度有固有的层次结构吗？

② 做出业务决策仅仅需要当地的有关信息吗？

③ 某些产品是否仅仅在某一地区销售？

（4）关键性能指标：不同的用户会有不同的看法。例如，部门的绩效是怎样监测的？部门内部提供哪些关键的指标？

（5）信息频率：可以从用户处理信息的时间灵敏度获得信息频率。如用户需要多长时间对数据更新一次？适当的时间结构是什么？在数据仓库中，对信息有实时性需求吗？

2. 对用户需求调研结果的分析

在与用户交流阶段，应该确定数据仓库需要访问的有关信息。用户应该清晰地确定所需的信息。例如，数据仓库的用户需要得到有关产品收入的详细统计信息，包括过去 5 年中的年龄、组别、性别、位置和经济状况等信息。

然后根据用户的信息需求，抽取出信息的度量值和维度信息，例如对于需要观察的产品收入，可以确定其度量指标和维度如下。

（1）度量指标：包括产品销售的实际收入、产品销售的预算收入及产品销售的估计收入。

（2）维度：包括已经销售的产品信息、销售地点（位置信息）和顾客信息（如年龄组别、性别、位置和经济状况）等。

假定 Adventure Works 的销售和营销团队以及高级管理人员对数据分析有如下需求。

（1）目前的报表是静态的。用户无法通过交互方式探测报表中的数据以获取更详细的信息，例如，他们可以处理 Microsoft Office Excel 透视表。虽然现有的一组预定义报表足以供许多用户使用，但更高级的用户却需要对数据库进行直接查询访问，以进行交互式查询和访问专用报表。

（2）查询性能差异很大。例如，有些查询只需几秒钟便可返回结果，而另一些查询需要几分钟才能返回结果。

（3）用户所在的业务部门不同，其感兴趣的数据视图也不同。每个组都很难理解与其不相关的数据元素。

（4）业务用户很难构造一些专用查询，以组合两个相关的信息集（如销售额和销售配额）。此类查询会占用大量的数据库空间，因此，公司要求用户向数据仓库团队请求跨主题区域的数据集。

（5）希望通过一个通用的元数据层提供统一的数据访问以进行分析和报告。

（6）简化用户的数据视图，从而加速交互式查询、预定义查询以及预定义报表的开发。

总之，通过对历史数据和需求的分析，可以明确用户正在使用的数据现状、他们如何使用这些数据及他们将利用数据仓库干什么。充分的交流为数据仓库的总体设计打下基础。

3.2.3 采用信息包图法设计数据仓库的概念模型

在收集分析需求并作了详细的需求调研之后，我们对企业需求有了一个比较清晰的了解，这时可以对数据仓库数据库的概念模型作设计，通常采用面向主题的自顶而下的设计方法，数据仓库的概念模型将面向主题，也就是面向对象，示例数据库中的对象如客户、产品和供应商等多维信息。终端用户通过各种维度来获取业务数据，其中时间是最基本、最关键的维度。对于面向主题的数据仓库同传统的数据库设计一样需要经历概念模型设计、逻辑模型设计和物理模型设计三个阶段(如表 3.4 所示)。与之相对应，数据仓库的设计方法分别是针对数据仓库的信息包图设计、星型图模型设计和物理数据模型设计，要求如下。

(1) 数据仓库的概念模型通常采用信息包图法来进行设计，要求将信息包图的 5 个组成部分(名称、维度、类别、层次和度量)全面地描述出来。

(2) 数据仓库的逻辑模型通常采用星型图法来进行设计，要求将星型图的 5 类逻辑实体(度量逻辑实体、维度逻辑实体、层次逻辑实体、详细信息逻辑实体和类别逻辑实体)完整地描述出来。

(3) 数据仓库的物理模型通常采用物理数据模型法来进行设计，要求将物理数据模型的 5 类表(事实表、维表、层次表、详细信息表和类别表)详细地描述出来。

表 3.4 数据仓库与 OLTP 数据库的设计方法比较

设计阶段	数据仓库	OLTP 数据库
概念模型	信息包图	数据流程图
逻辑模型	星型图模型	实体关系图
物理模型	物理数据模型	物理数据模型

在与用户交流的过程中，上两个节确定了数据仓库所需要访问的信息，这些信息包括当前的、将来的以及与历史相关的数据。本节将确认操作数据、数据源以及一些附加数据需求，建立信息包图，进而确定数据仓库中的主题和元数据，有效地完成查询和数据之间的映射，完成概念数据模型的设计。

1. 信息包图法简介

由于数据仓库的多维性，利用传统的数据流图进行需求分析已不能满足需要。因此，数据仓库的建模包括超立方体(hypercube)法及信息包图法。

超立方体法也是采用自上而下的方法设计，其步骤如下。

(1) 确定模型中需要抓住的业务过程，例如销售活动或销售过程。

(2) 确定需要捕获的度量值，例如销售数量或成本。

(3) 确定数据的粒度，即需要捕获的最低一级的详细信息。

由于超立方体法在表现上缺乏直观性，尤其是当维度超出三维后，数据的采集和表示都比较困难，这时可以采用信息包图法在平面上展开超立方体，即用二维表格反映多维特征。信息包图提供了一个多维空间来建立信息模型，并且提供了超立方体的可视化表示。

信息包图定义主题内容和主要性能指标之间的关系，其目标就是在概念层满足用户需求。信息包图拥有三个重要对象：(度量)指标、维度和类别。利用信息包图设计概念模型