浅析数据仓库技术及其实现

胡光正 （北方交通大学自动化系统研究所 100044）

摘要：本文在简要论述数据仓库技术产生背景的基础上，对数据仓库技术进行了介绍，并分析了实现这一技术的方案，希望对将要建立数据仓库的企业有所借鉴，使数据仓库能够为企业提高经济效益起到良好的促进作用。

关键词：数据仓库 数据集市

一、前言

随着全球范围内市场竞争的日益激烈，越来越多的企业将视线转移到企业信息库存，以便更好地了解客户和业务运作。他们希望将来自不同数据源的业务数据进行提练和综合，为企业提供准确的决策支持依据，发现庞大信息库存中隐藏的财富，并且赶走他们的竞争对手之前将这些信息转化为资产，从而在竞争中立于不败之地。

数据仓库技术也就在此时应运而生。

这一技术具有十分强大的功能。例如，它可以帮助企业真正了解客户的需求，他们在利用何种软件以及产生了那些不必要的开销，客户的支付方式以及支付周期等等。通过数据仓库技术的帮助，企业可以向用户提供他们所期待的个性化的产品与服务，这将大大提高企业与用户之间的相互信任。因此可以预见，这一技术的应用范围越来越广，其发展前途是令人乐观的。

二、数据仓库

数据仓库技术通俗来讲，就是一种数据组织的高级形式。在当今的信息社会，数据组织技术的发展大致经历了五个阶段：

人工管理阶段(50 年代)：组织数据进行科学计算；

文件系统阶段(60 年代)：进行相对简单的数据处理；

集中式数据库阶段和分布式数据库阶段(70～80 年代)：进行实时联机和相对复杂的数据处理；

数据仓库和数据集市阶段(90 年代初至今)：除了进行前者的工作外，还要进行高层次的数据分析。

从上面的事实我们看出，数据组织技术的发展异常迅猛，愈加满足了企业多方面应用需求。

一个现代化的企业对数据的运用可以概括为图 1 的层次。

图 1 中，数据在不同层次具有不同的意义：为基层人员提供基本业务信息；为中层干部提供管理信息；为高层人员提供决策信息。以前，数据组织技术一般为下层人员提供服务，无法进行高层决策。实现数据仓库技术的目的就是要填补这一空白。

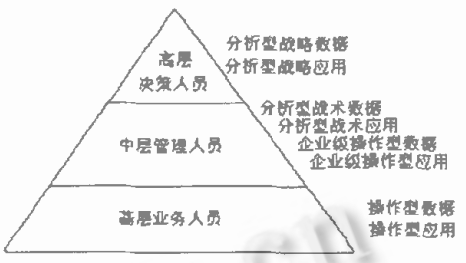

图 1 企业模型

本世纪八十年代中期，Bill Inmon 先生首先提出了“数据仓库”(Data Warehouse)这一名词。它最初被设计成一个商业数据库，具有稳定性(主要成分不变)、历史性(包含历史信息)和面向主题性(由客户、产品和市场组成)。这些最初的“数据仓库”根据客户产品销售情况和财务状况等信息的分析，得出对企业活动的整体认识。在开始阶段数据仓库的建立可以概括为四个步骤：

(1)设计出一个包含商业数据和信息的数据库，为商业实体所用；

(2)开发收据抽取和转换程序，从产品系统将数据取出后放入数据仓库；

(3)开发数据仓库得到实时的更新；

(4)购置查询和报表生成工具，使用户通过企业内部

网和个人计算机极为方便地获取信息。

Inmon 先生的这一思想对众多企业确实产生了很大的吸引力。可是开始在实际操作中却遇到了不少麻烦。尽管不少商人为建立本企业的数据仓库投入了大量的人力和物力，但是由于技术本身的不成熟，并未取得应得的回报，开发进度一再延长，仍然无法让用户给出明确的需求定义，使用户和开发人员都处于不好处理的境地。

近两年来，针对上述问题，出现了一种新的解决方法，这就是数据集市技术。数据集市也是一种数据仓库，只是它更精炼、更加面向主题。数据集市的优势在于开发周期的缩短和费用的大幅度降低。由于企业的数据庞大，真正将这些数据集中在一个数据库中几乎是不可能的。运用数据集市技术以后，设计、抽取、转换、加载和查询变得更加简便，用户中的一部分人能够更加精确地知道他们所需要的信息。

Inmon 先生又给出了一个更为精确的定义：数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合。

专业人员将这一技术概括为以下四点：(1)它是一个处理过程，而不仅是代表一组产品；(2)它是从大量的企业数据中发现有价信息的过程；(3)它可以充分利用现有资源，而不是摈弃重构；(4)它提供了系统数据的多种访问形式。

数据仓库的体系结构可以概括为图 2。

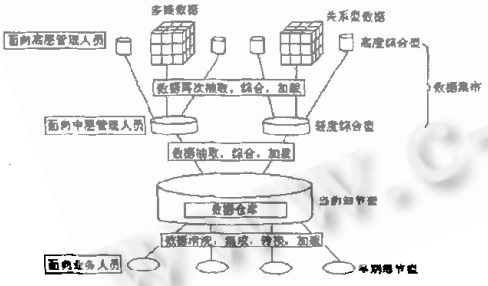

图 2　数据仓库(多层)体系结构图

三、数据仓库的建立过程

1.准备阶段

在这里必须确立企业所采用的数据仓库模型结构。目前大多数企业一般采用以下三种体系结构：

(1)集中式数据仓库模型(图 3)，它的特点是：

①数据能不断地从数据源系统累计到数据库中；

②数据的存放与取舍规则和 OLTP(可操作数据源)系统独立；

③数据库中存放的是企业的数据，可以跨业务领域；

④OLTP 的数据相互独立，它们的性能不受影响。

(2)Inmon 先生提出的分布式体系结构(见图 4)：如果数据源的存储形式不相同，需要 ODS (Operational Data Store)使数据格式以统一的形式存储并且加以归纳。如果各个数据源格式相同，则可以不设 ODS，即数据直接被数据仓库和数据集市使用。

(3)集中式数据仓库模型(见图 5)：这种模型除了具有第一种模型的特征外，其数据更加靠近最终用户。上述三种模型各具特色，企业在建模时可以根据自身的情况进行选择，以期得到最佳回报。

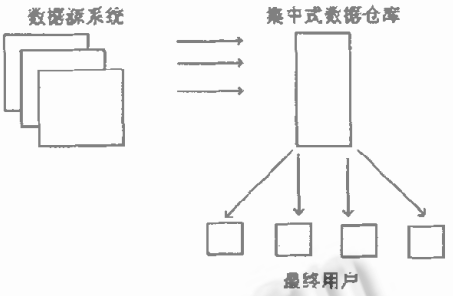

图 3　集中式数据仓库模型

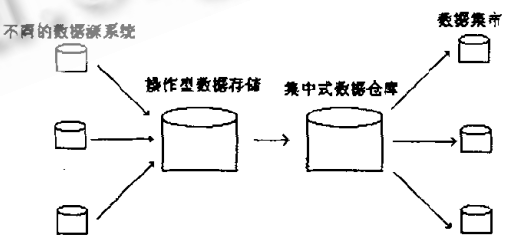

图 4　Inmon 先生提出的数据仓库模型

2.设计阶段的工作

(1)获取最终用户的业务需求；

(2)定义业务规则；

(3)建立主题设计视图；

(4)建立企业逻辑数据模型;

(5)定义操作型数据源;

(6)定义相应数据仓库的数据模型;

(7)必要的话,将数据仓库的模型分解成几个子模型分别建设,以满足不同用户和不同工具的需要。上述每一步在执行中都要经过严格的质量控制,因此每一步的实施可能要经过多次反复。

在这个阶段,主题视图的合理转换是最为重要的问题。

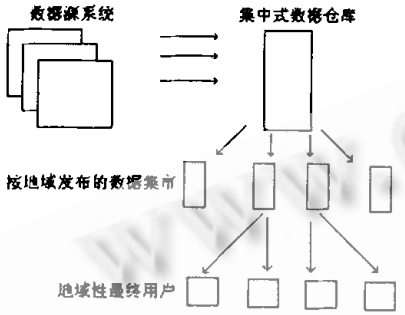

图 5　集中分布式数据仓库模型

主题视图向逻辑数据模型的转换和将逻辑数据模型转化为数据仓库数据模型可以按照图 6 的模式进行。

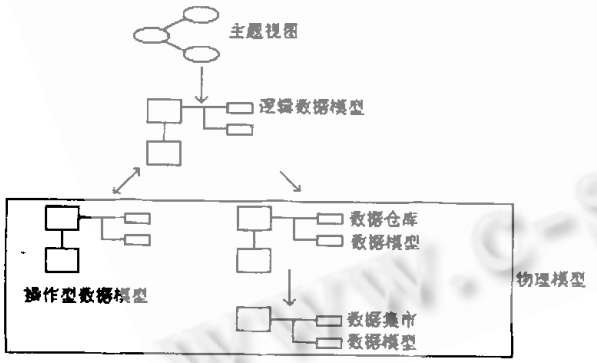

图 6　模型的转换

我们应当充分理解两者的区别:(见下表)

逻辑数据模型	数据仓库数据模型
范式化	非范式化
详细数据	详细和汇总数据
企业运作角度	企业决策和战略角度
没有派生数据	含有派生的战略数据
无数据数组	有数据数组
以企业规划为中心	以数据的使用和稳定性为中心

根据上表,在进行数据转换时可以遵循以下原则:

(1)删除非战略性数据,这里主要指与决策支持系统无关的数据;

(2)增加时间主键;

(3)增加合并数据项;

(4)加入不同级别粒度的汇总数据;

(5)加入数据的汇总模式;

(6)将不同表中的数据合并;

(7)根据数据的稳定性进行数据的分离处理。

建立物理模型的最终目的就是要建立数据集市,为企业决策提供依据,数据集市是从数据仓库中派生出来的,可以与数据仓库处于同一平台,也可以分开。同时我们还必须明确它是数据仓库的一部分。目前主要有三种模型可以实现数据集市:多维型、关系型和非关系型。其中多维型的应用最为广泛。传统的数据模型比较复杂,最终用户难于理解。多表连接查询既费时,又占用大量资源。多维模型可以解决这个问题,它是人们观察数据的形象表示,可以是二、三、四甚至更多维。我们可以对多维模型进行切片或切块(选择哪维或哪几维作为查询条件),维的增加可以使事实表的数据更加细节化。需要指出的是,事实表数据的过于细节化,会造成数据模型的失真,最终甚至使宏观模型失去本身的意义,产生过犹不及"的效果。

多维模型至今已经出现了四种形式:

(1)星型模式:用一个事实表对应多个组件;

(2)雪花模式:为了避免数据冗余,用多张表描述一个复杂维,在星型模式的基础上,构造出表的多维结构,实现了对维表的再次分割,对维的属性进行了多层次分类;

(3)星座模式:设置了多个事实表,它们之间由维表建立间接的联系,其余特点和星型模式相同;

(4)雪暴模式:是四种形式中最为复杂的一种,几乎包容了上述三者的特点,具有多个事实表和多层维表。企业可以根据自身业务的复杂程度选取适当的方案。

3.物理实现阶段

这一阶段的流程可以用图 7 表示。

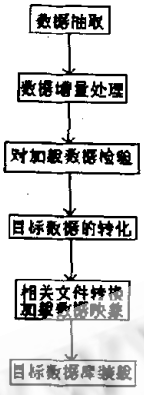

图 7　数据集成与数据转换过程

这一阶段的关键是制定数据集成、抽取、清洗和转换的时机,根据现场专业人员的实际经验,主要确定下列时机:何时抽取现有的传统数据;何时处理现有的传统数据;何时集成和验证来路不明的传统数据;何时汇总数据;何时往数据仓库中加载数据;合适备份过时数据和何时清除不用的数据;何时加载轻度汇总数据;何时建立汇总数据的索引;何时建立轻度汇总数据来产生高度汇总数据;何时加载高度汇总数据;何时建立高度汇总数据的索引。

4.应用阶段

这一阶段的主要任务是建立数据分析。

现在可供选择的数据分析模型有四种:

绝对模型用于静态数值分析,通过比较历史数据和行为描述过去发生的事情;解释模型用于静态数值分析,通过层层细化,找出事实发生的原因;思考模型用于动态数值分析,它通过引入一定的参数后,预测将来发生的事情;公式模型用于最高级动态数据分析,知道需要引入那些参数以及所产生的后果。

数据分析的工具有三种类型:查询工具:指对分析结果的查询,而不是对记录级的查询;验证型工具:从数据仓库中发现事实,实现数据分析的前三种模型;挖掘型工具:从大量数据中发现模式,实现第四种分析模型。

前两种工具已应用较为普遍。目前这一领域最为热门的研究项目集中于实现数据挖掘。这一工具的实现可以不基于所建立的数据仓库,因此具有很大的灵活性。数据挖掘通过依次完成数据准备(数据集成、数据选择和预分析)、数据挖掘(关联分析、序列模式分析、分类分析和聚类分析)、数据表达(以直观的、便于用户理解和观察的方式表达)和评价(如果对上述过程的结果不满意,可以重复上述过程,直到满意为止)来实现。

5.运行与维护阶段

简单说,主要任务是维护数据仓库的不断变化。

通过上述阶段基本可以完成数据仓库的建设过程。

四、结束语

目前,在众多软件厂商大力开拓数据仓库技术的带动下,基于这一技术的解决方案正在趋向成熟,在集成、可扩展性、安全性、易管理性、数据挖掘性能、数据集市同步性等方面都有相当大的提高。在他们的带动下,许多合理的解决方案纷纷出台。数据仓库技术已经广泛应用于电信行业(成本管理)、金融服务(有价证券风险管理和财政预算计划)、信用卡运作、医疗保健、保险理陪等众多领域。事实证明,由于数据仓库技术真正解决了困扰客户的业务问题,在很大程度上提高了企业的经济效益,因此使众多用户增强了对这一技术的信息,看到了数据仓库技术为他们带来的新契机。虽然这一技术的发展时间还很短,但是它当前已经取得的成就使人们相信,随着逐步成熟,它将在企业发挥更加重要的作用。

参考文献

[1] SYBASE 软件(北京)有限公司市场部编《SYBASE 世界》1998 年第 2 期

[2] 刘韩"BataBlade 与数据仓库"《计算机世界报》1998 年 6 月 2 日

[3] 刘韩 李延辉"数据仓库动力引擎"《计算机世界报》1998 年 3 月 23 日

(来稿时间:1998 年 6 月)