第 1 章　多维数据分析技术概述

1.1　多维数据分析技术概念

多维数据分析也称为联机分析处理(on-line analytical processing,OLAP)是以海量数据为基础的复杂分析技术。它支持各级管理决策人员从不同的角度,快速灵活地对数据库中的数据进行多角度查询和分析,并以直观易懂的形式将查询和分析结果展示给决策人员。

在实际分析过程中,用户需要的数据往往不是某一指标单一的值。他们希望能从多个角度观察某一指标或多个指标的值,并且找出这些指标之间的关系。比如,决策者可能想知道"东部地区和西部地区今年 6 月份和去年 6 月份在销售总额上的对比情况,并且销售额按 10 万～20 万、20 万～30 万、30 万～40 万以及 40 万以上分组"。上面的问题是比较有代表性的。决策所需数据总是与一些统计指标(如销售总额)、观察角度(如销售区域、时间)和不同级别(如地区、统计值区间划分)的统计(或合并)有关,我们将这些观察数据的角度称为维。可以说决策数据是多维数据,多维数据分析是决策的主要内容。但传统的关系数据库系统及查询工具对于管理和应用这样复杂的数据显得力不从心。

多维数据分析是专门为支持复杂的分析操作而设计的,侧重于决策人员和高层管理人员的决策支持,可以应分析人员的要求快速、灵活地进行大数据量的复杂处理,并且以一种直观易懂的形式将查询结果提供给决策人员,以便他们准确掌握企业(公司)的经营状况,了解市场需求,制定正确方案,增加效益。

多维数据分析是以数据库或数据仓库为基础的,其最终数据来源与联机事务处理(on-line transaction processing,OLTP)一样均来自底层的数据库系统,但二者面对的用户不同,数据的特点与处理也明显不同。

多维数据分析与 OLTP 是两类不同的应用。OLTP 面对的是操作人员和低层管理人员,多维数据分析面对的是决策人员和高层管理人员。OLTP 是对基本数据的查询和增删改操作处理。它以数据库为基础,而多维数据分析更适合以数据仓库为基础的数据分析处理。多维数据分析所依赖的历史的、导出的及经综合提炼的数据均来自 OLTP 数据库。多维分析数据较之 OLTP 数据要多一步数据多维化或综合处理的操作。例如,对一些统计数据,首先进行预综合处理,建立不同级别的统计数据,从而满足快速统计分析和查询的要求。除了数据及处理上的不同之外,多维数据分析的前端产品和界面风格及数据访问方式也同 OLTP 有别。多维数据分析多采用便于非数据处理专业人员理解的方式(如多维报表、统计图形)显示数据,用户可以方便地进行逐层细化及切片、切块、旋转

等操作，而 OLTP 的数据显示比较固定和规范（二维表）。

多维数据分析中包括如下一些基本概念。

1. 度量属性

度量属性（measure）是决策者所关心的具有实际意义的数量，例如，销售量、库存量等。度量属性所在的表称为事实数据表，事实数据表中存放的事实数据通常包含大量的行。事实数据表的主要特点是包含数值数据（事实），而这些数值数据可以汇总以提供有关单位运作历史的信息。每个事实数据表包括一个或多个列，这些列作为相关的维度表引用的外码。事实数据表一般不包含描述性信息，也不包含数字度量字段以及使事实数据表和维度表之间相关联的字段之外的任何数据。

在多维数据集中，通常对基于该多维数据集的事实数据表中某个列或某些列的值，进行聚合和分析，这些值就称为度量值。度量值是多维数据集中的一组值，这些值基于多维数据集的事实数据表中的一列或多列，而且通常是数值。度量值是所分析的多维数据集的中心值，它是最终用户浏览多维数据集时重点查看的数值数据。用户所选择的度量值取决于最终用户所请求的信息类型。

2. 维度

维度（dimension）是人们观察数据的特定角度，简称为维。例如，企业常常关心产品销售数据随时间推移而产生的变化情况，这是从时间的角度来观察产品的销售，因此时间就是一个维（时间维）。企业也时常关心自己的产品在不同地区的销售分布情况，这是从地理分布的角度来观察产品的销售，所以地理分布也是一个维（地理维）。

包含维度信息的表是维度表，维度表包含描述事实数据表中的事实记录的特性。有些特性提供描述性信息；有些特性则用于指定如何汇总事实数据表数据以便为分析者提供有用的信息。

3. 维的层次

人们观察数据的某个特定角度（即某个维）还可以存在细节程度不同的多个描述方面，我们称这多个描述方面为维的层次（dimension hierarchy）。一个维往往具有多个层次。例如描述时间维时，可以从日、月、季度、年等不同层次来描述，那么日、月、季度、年等就是时间维的层次；同样，城市、地区、国家就构成了地理维的多个层次。

4. 维度成员

维的一个取值称为该维的一个维度成员（dimension member），简称为维成员。如果一个维是多层次的，那么该维的维度成员是在不同维层次的取值的组合。例如，我们考虑时间维具有日、月、年这 3 个层次，分别在日、月、年上各取一个值组合起来，就得到了时间维的一个维成员，即“某年某月某日”。一个维成员并不一定在每个维层次上都要取值，例如“某年某月”、“某月某日”、“某年”等都是时间维的维成员。

图 1-1 显示了在多维数据集中的这些概念。

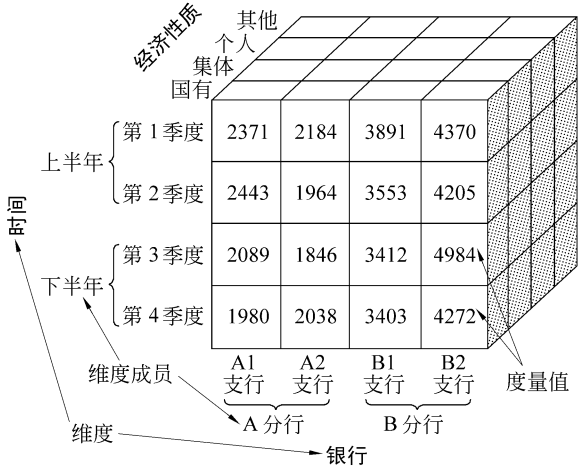

图 1-1　多维数据分析示例

1.2　多维数据分析方法

多维分析可以对以多维形式组织起来的数据进行切片、切块、旋转等各种分析操作，以便剖析数据，使分析者、决策者能从多个角度、多个侧面观察数据库中的数据，从而深入了解包含在数据中的信息和内涵。多维分析方式迎合了人的思维模式，减少了混淆，并降低了出现错误解释的可能性。

多维数据分析通常包括以下几种分析方法。

1. 上卷

上卷(roll-up)是在数据立方体中执行聚集操作，通过在维层次中上升或通过消除某个或某些维来观察更概括的数据。例如，图 1-2 所示的数据立方体(水平轴为商品类别维，垂直轴为时间维，Z 轴为地点维)经过沿着地点维的概念层次上卷，由城市上升到国家，得到图 1-3 所示的立方体。现在销售数据不是按照城市分组求值，而是按照国家分组求值了。

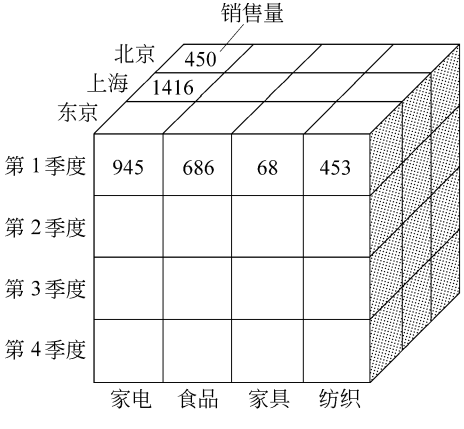

图 1-2　多维数据立方体

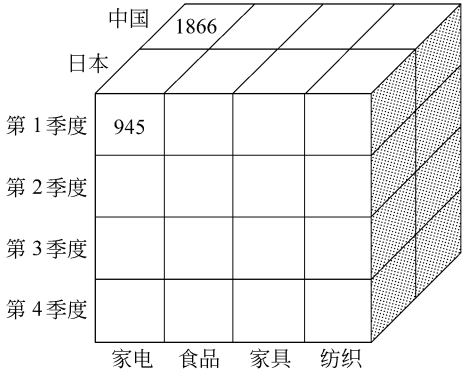

图 1-3　图 1-2 上卷后的效果

也可以通过消除一个或多个维来观察更加概括的数据。例如，图 1-4 所示的二维立方体就是通过从图 1-2 所示的三维立方体中消除了“国家”维后得到的结果，这里所有国家的销售数据都累计在一起了。

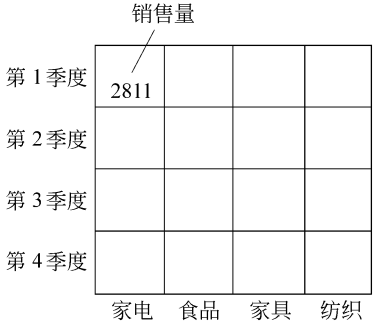

图 1-4　图 1-2 消除“国家”维后的结果

2. 下钻

下钻(drill-down)是通过在维层次中下降或通过引入某个或某些维来更细致地观察数据。

例如，对图 1-2 所示的数据立方体经过沿时间维进行下钻，由季度下降到月，就得到了如图 1-5 所示的数据立方体。现在的销售数量不是按季度计算，而是按月进行计算了。

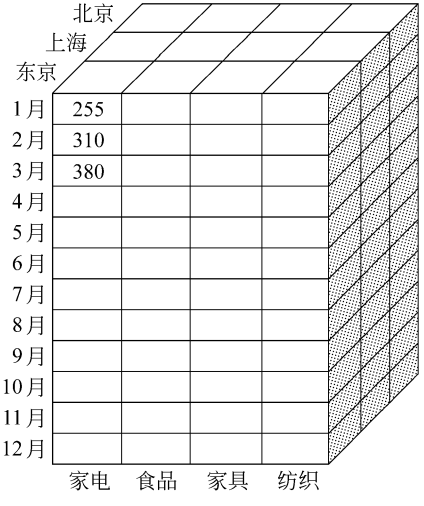

图 1-5　图 1-2 的下钻结果

3. 切片

在给定的数据立方体的一个维上进行的选择操作就是切片(slice)。切片的结果是得到一个二维的平面数据。

例如，在图 1-2 所示数据立方体上，使用条件“时间＝第 1 季度”进行选择，就相当于在原来的立方体中切出一片，结果如图 1-6 所示。

4. 切块

在给定的数据立方体的两个或多个维上进行的选择操作就是切块(dice)。切块的结

果是得到一个子立方体。

例如，在图 1-2 所示的数据立方体上，使用条件：

（地点＝“北京” or “上海”）

And（时间＝“第 1 季度” or “第 2 季度”）

And（商品类型＝“家电” or “食品”）

进行选择，相当于在原立方体中切出一小块，结果如图 1-7 所示。

图 1-6　图 1-2 的切片结果

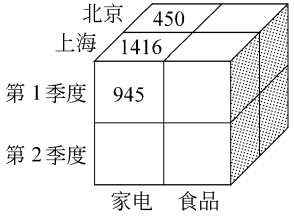

图 1-7　图 1-2 的切块结果

5. 转轴

转轴（pivot or rotate）就是改变维的方向，将一个三维立方体转变为一系列二维平面。

例如，图 1-8 所示是图 1-6 的二维切片的“商品轴”和“地点轴”交换位置的结果。

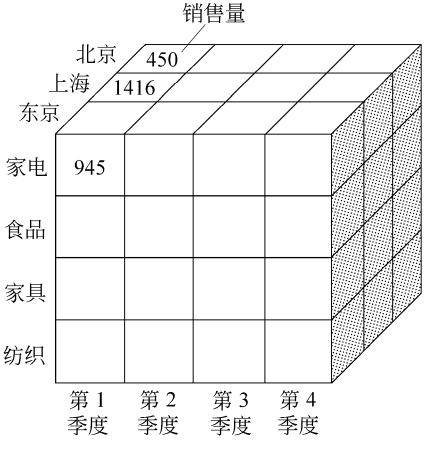

图 1-8　图 1-6 转轴后的结果

1.3　事实数据与维度数据的比较与识别

我们设计一个数据仓库架构首先要面对的就是决定哪些是事实数据，哪些属于维度数据。我们要从众多字段中决定事实与维度数据。这是一件麻烦的事，但是绝对不能轻视这项工作，因为事实表结构的正确与否决定了整个数据仓库计划的成败。表 1-1 显示了事实数据与维度数据的区别。

表 1-1　事实数据与维度数据的特性

项　目	事 实 数 据	维 度 数 据
规模	几百万笔/上亿笔数据	远比事实数据少
数据标识	拥有多个外键	拥有单一主键
数据类型	数值数据	文字语句数据
数据性质	不会更新	经常改变

若数据不会随着时间而改变，则它很可能是事实数据。若发觉可以使用一项数据找到很多笔记录，则该项数据很有可能是维度数据。一般来讲，在一个数据仓库中肯定有时间数据，而且时间不会是事实数据。

识别事实数据和维度数据分四个步骤：

- 搜索最基本交易，它们极可能是事实数据。
- 决定搜索每一事实数据的键，它们极可能是维度数据。
- 检验每一可能是事实数据的字段，确定它不是嵌在事实数据中的维度数据。
- 检验每一可能是维度数据的字段，确定它不是嵌在维度数据中的事实数据。

当发现一些事实数据可能是维度数据时，将它置入维度数据中，当发现一些维度数据可能是事实数据时，将它置入事实数据中。该过程包含了两个循环，我们要不停地检验，一直到确定所有的事实与维度数据为止。

1.4　审计实务中应用多维数据分析技术的重要意义

为了顺应现代科学技术发展的潮流，审计署从 20 世纪 90 年代末开始大力推广计算机技术在审计实务中的应用。从 Excel 表格到 SQL Server 等大型数据库软件的应用，再到审计软件的开发，我们正在计算机审计的道路上不断前进，审计人员的计算机应用水平有了长足的提高。关系数据库技术已在政府审计中得到了广泛的应用，许多审计人员已经能够灵活地在审计过程中编写 SQL 语句。但是，目前无论是利用已有的商业软件还是自身开发的应用软件，都是在审计过程中的一个局部应用，很多时候还是脱离不了利用计算机开展“辅助审计”的影子，盲目性大，审计风险较高，还没有把计算机审计作为一种崭新的思维方式和审计方式贯穿到实务中。在实际工作中，我们往往缺乏一种把握总体、统领全局的能力和技术。财政审计、税收审计、海关审计、金融审计、企业审计数据量庞大、数据表复杂，如何在这个数据迷宫中迅速找出审计所需要的信息，是一个不得不面对的难题。这个问题不解决，就不可能彻底杜绝盲人摸象的现象。

如何在审计过程中迅速把握总体，如何从被审计单位浩如烟海的电子数据中根据需要找出有用的信息成了摆在我们面前的迫切需要解决的问题。审计人员的总体分析要求对关系数据库进行大量计算才能得到结果，简单的 SQL 语句和小型数据库软件已经无法满足这样的要求了。我们需要利用专门的数据综合引擎和直观的数据访问界面，以统一复杂查询中多种多样的应用逻辑，使系统在很短的时间内响应审计人员的复杂查询。因此，我们提出了在现场审计过程中应用多维数据集和多维分析的问题。

构建总体分析模型，站在一定高度上把握总体，从观察趋势、选择重点，到运用钻取、掌握明细，直至发现线索、引导延伸，这就是多维分析的总体过程。这不仅是一种审计技术，更是一种审计的思维方式。作为一种崭新的审计方式，它带来了以下两点革命性的变化：一是从瞎子摸象转变为把握总体；二是从进点后摸线索转变为带着线索进点。

开展多维数据分析技术在审计中的应用研究，探索如何将这一全新的技术在实践中贯穿、应用并加以推广，已经成为提升计算机审计水平和层次的当务之急。多维分析是一种通用技术，要应用到审计实践中去，必须紧密结合审计的需求和特点，必须详细分析被审计电子数据的特点和规律，而不可能简单地照搬照抄，诸如对从被审计单位采集到的源数据如何清理、转换、验证，如何建立面向审计分析的审计中间表，如何根据审计需求设计和建立维度，如何分析已建立起来的多维数据集，这都需要一一加以深入研究。而这些也正是本书试图回答的问题。

第 2 章　多维数据分析工具

数据仓库技术越来越受到广泛的关注，越来越多的审计人员意识到建立 OLAP 所能带来的好处。利用 OLAP 进行数据分析，可以帮助审计人员从多个角度观察数据，甚至预测发展趋势。多维数据分析工具就是帮助审计人员进行多角度数据分析的能力强大的工具。

同一般的数据库管理系统一样，多维数据分析的工具也可以划分为客户端工具和服务器端工具。服务器端工具主要用于保存多维数据分析的聚合数据和元数据，客户端工具主要用于查询和显示多维数据的分析结果。

本章我们介绍一些比较常用的服务器端多维数据分析工具和客户端多维数据分析工具。

2.1　常用的服务器端分析工具

可以用作多维数据分析的服务器端工具很多，其中常用的、功能比较强大的有两个：一个是微软(Microsoft)公司的 SQL Server Analysis Services(SQL Server 分析服务)，另一个是 IBM 公司的 DB2 OLAP Server(OLAP 服务器)。下面对这两种服务器端工具作简要介绍。

2.1.1　Microsoft SQL Server Analysis Services

20 世纪 90 年代早期 OLAP 工具的最大问题是对最终用户来说不太好用，要由开发人员在这些工具之上再开发新的应用程序来减轻最终用户的负担。这里的最终用户是指分析数据仓库中积累的历史数据，并根据得到的结果进行决策的人。

微软公司意识到为数据仓库开发一个更好的系统的重要性，同时也意识到开发一种使分析过程轻松而愉快的分析工具的重要性。微软首先在 SQL Server 7.0 版本中提供了这个问题的解决方案，并在 SQL Server 2000 中进一步完善了这个方案，其中包括了新版的 OLAP Services，同时增加了数据挖掘的功能，并将它们称为 Analysis Services。

Analysis Services 提供了从数据仓库中设计、构建及管理多维数据集的能力，同时也可以让客户端取得 OLAP 数据。以下从分析服务的特点、体系结构、存储结构三个方面来介绍分析服务。

2.1.1.1　分析服务的特点

SQL Server 2000 的分析服务除了为客户端工具提供多维数据外，还具有创建和管理多个多维数据以及存储元数据的功能。分析服务的主要特点如下：

- 易用性：操作中的任何一步都有很多向导、编辑器和帮助材料提供帮助。用户通过分析管理器(Analysis Manager)提供的操作界面，可以方便地访问元数据和多维数据集，而且也可以使用向导建立和编辑多维数据集、维度和级别。
- 灵活的数据存储模型：Analysis Services 为维度、分区以及多维数据集提供了多种存储模式。可以将多维数据集存放在多维立方文件(MOLAP)或者关系型数据库中，或是这两种的混合。多维数据集还可以被分区，并且以不同的模式存放分区。多维数据集的存储将在本节后面作详细介绍。
- 伸缩性：Analysis Services 同时支持基于 Intel 的服务器和 DEL Alpha 服务器。OLAP 客户端可以在 Windows 9x、Windows NT 和 Windows 2000 平台上运行。Analysis Services 还解决了很多数据仓库中的问题，比如：自定义聚集选项、基于应用的优化、数据压缩以及分布计算等。这一切使 Analysis Services 具有很强的伸缩性。
- 集成：Analysis Services 与微软管理控制台(Microsoft Management Console，MMC)集成在一起，可以将 Analysis Services 作为 MMC 的一个部件。Analysis Services 的安全性也集成在 SQL Server 和 Windows NT 的安全机制中。
- 支持大量的 API 和函数：OLAP 服务器和数据透视表服务(Pivot Table)支持 OLE DB、ADO、用户自定义函数和决策支持对象(Decision Support Object，DSO)。
- 分布式处理能力：通过分区(partition)不仅可以调整多维数据集的大小，还可以把多维数据集分布在多个服务器上，以便并行处理。
- 服务器端结构的高速缓存：对于服务器端，可以利用分析服务的高速缓存来查询多维数据以及元数据。这样就可以根据内存中的数据查询，而不用访问磁盘上的数据，从而减轻网络流量，并加快查询的响应速度。

2.1.1.2 分析服务的体系结构

SQL Server 2000 Analysis Services 提供了对数据仓库数据的快速访问。通过在多维结构中对数据仓库中的数据进行提取、汇总、组织和存储，可以对最终用户查询做出快速响应。

图 2-1 显示了 Analysis Services 的系统体系结构。从图中，可以看出 Analysis Services 的体系结构包含三个层次：操作的数据源、Analysis Services 及其工具、提供报表和其他商务智能服务的客户端应用。其中中间部分的分析服务又由几个高层组件构成，这些组件包括数据转换、数据存储、Analysis 服务器和数据透视表服务。数据存储包括数据仓库和数据集市、OLAP 数据库和挖掘模型。

(1) 数据源

Analysis Services 和数据仓库(包括数据集市)的数据源是由向数据仓库提供数据的操作型数据和一些服务组成。这些服务将数据从源转换成可以存储在数据仓库、数据集市或普通关系型数据库的格式，最后存入 OLAP 多维数据集。

在 Analysis Services 中，操作型数据通常是一个关系型数据库。但任何可通过

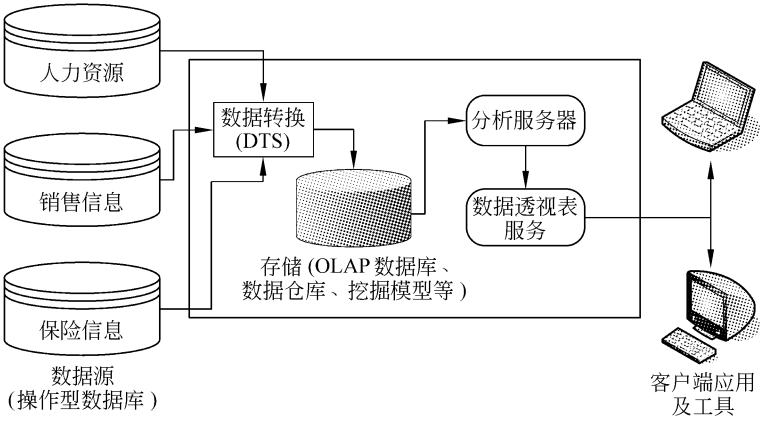

图 2-1　Analysis Services 体系结构

ODBC 或 OLE DB 接口连接的数据都可以作为 Analysis Services 的数据源。微软的 DTS 工具可以将数据从这些源转换到 Analysis Services 中。

(2) 数据转换服务

微软的数据转换服务(Data Transformation Services,DTS)支持数据的导入、导出和转换,它完成数据从 OLTP 源转移到 OLAP 系统的工作,在转换过程中,DTS 可以进行数据校验、清理、合并和必要的转换工作。

(3) 数据透视表服务

数据透视表服务支持用户通过定义要聚集的行和列即时创建交叉表。同时微软对数据透视表服务也做了一些改进,以适应 SQL Server 2000 中的数据挖掘和分析的需求,增强分析服务器、多维数据集间的通信。

在 OLAP 客户端,Analysis Services 可以和多个工具协同工作,如 English Query、Microsoft Office(特别是 Excel 和 Access),这些工具通过数据透视表服务可访问多维数据。由于一般用户对 Excel 和 Access 工具比较熟悉,因此,就可以更好、更方便地进行数据分析。

下面我们分别介绍分析服务的服务器端和客户端的体系结构。

1. 服务器端体系结构

Analysis Services 提供服务器功能以创建和管理 OLAP 多维数据集及数据挖掘模型,并通过透视表服务为客户端提供数据。服务器端操作通常包括:

- 从关系数据库,通常是数据仓库,创建并处理多维数据集。
- 以多维结构、关系数据库或二者的结合形式存储多维数据集数据。
- 从多维数据集或关系数据库创建数据挖掘模型,通常是在数据仓库中创建。
- 以多维结构、关系数据库或标准化 XML 格式的预测模型标记语言(PMML)的形式存储数据挖掘模型的数据。

图 2-2 显示了 Analysis Services 的服务器端体系结构。从该图中可以看到 Analysis Services 的组件及其相互关系。

图 2-2 所示体系结构的核心是 Analysis 服务器。安装 Analysis Services 时实际上是