

第3章 联机分析处理

在数据仓库系统中,联机分析处理(OLAP)是重要的数据分析工具。OLAP 的基本思想是企业的决策者应能灵活地,从多方面和多角度以多维的形式来观察企业的状态和了解企业的变化。

3.1 OLAP 概念

在信息爆炸的时代,信息过量几乎成为人人都需要面对的问题。如何才能不被信息的汪洋大海所淹没,从中及时发现有用的知识或者规律,提高信息利用率呢?要想使数据真正成为一个决策资源,只有充分利用它为一个组织的业务决策和战略发展服务才行,否则大量的数据可能会成为包袱,甚至成为垃圾。OLAP 是解决这类问题最有力的工具之一。

OLAP 专门设计用于支持复杂的分析操作,侧重对分析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查询处理,并且以一种直观易懂的形式将查询结果提供给决策制定者,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。OLAP 软件以它先进的分析功能和以多维形式提供数据的能力,正作为一种支持企业关键商业决策的解决方案而迅速崛起。

3.1.1 OLAP 的定义

在决策活动中,决策人员需要的数据往往不是单一指标的单一的值,他们希望能够从多个角度观察某个指标或者某个值,或者找出这些指标之间的关系。比如,决策者可能想知道“东北地区 and 西南地区今年一季度和去年一季度在销售总额上的对比情况,并且销售额按 10 万~50 万、50 万~100 万,以及 100 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)和不同级别的统计有关,可以将这些观察数据的角度称之为维。可以说决策数据是多维数据,多维数据分析是决策分析的主要内容。但传统的关系数据库系统及其查询工具对于管理和应用这样复杂的数据显得力不从心。

OLAP 是在 OLTP 的基础上发展起来的,OLTP 是以数据库为基础的,面对的是操作人员和低层管理人员,对基本数据的查询和增、删、改等进行处理。而 OLAP 是以数据仓库为基础的数据分析处理。它有两个特点:一是在线性(On Line),体现为对用户请求的快速响应和交互式操作,它的实现是由客户机/服务器这种体系结构在网络环境上完成的;二是多维分析(Multi-dimension Analysis),这也是 OLAP 的核心所在。

OLAP 超越了一般查询和报表的功能,它是建立在一般事务操作之上的另外一种逻辑步骤,因此,它的决策支持能力更强。在多维数据环境中,OLAP 为终端用户提供了复杂的数据分析功能。通过 OLAP,高层管理人员能够通过浏览、分析数据去发现数据的变化趋

势、特征以及一些潜在的信息,从而更好地帮助他们了解商业活动的变化。目前,普遍为人们所接受的 OLAP 的定义有两种。

1. OLAP 理事会给出的定义

联机分析处理(OLAP)是一种软件技术,它使分析人员能够迅速、一致、交互地从各个方面观察信息,以达到深入理解数据的目的。这些信息是从原始数据转换过来的,按照用户的理解,它反映了企业真实的方方面面。

企业的用户对企业的观察自然是多维的。例如销售,不仅可从生产这方面看,还与地点、时间等有关,这就是为什么要求 OLAP 模型是多维的原因。这种多维用户视图通过一种更为直观的分析模型进行设计和分析。

OLAP 的大部分策略都是将关系型的或普通的数据进行多维数据存储,以便于进行分析,从而达到联机分析处理的目的。这种多维数据库也被看做超立方体,沿着多个维度存储数据,为用户沿事物的任意的多个维度方便地分析数据。

2. OLAP 简单定义

近来,随着人们对 OLAP 理解的不断深入,有些学者提出了更为简明的定义,即联机分析处理是共享多维信息的快速分析(Fast Analysis of Shared Multidimensional Information),它体现了四个特征:

(1) 快速性(fast): 用户对 OLAP 的快速反应能力有很高的要求。系统应能在 5s 内对用户的大部分分析要求做出反应,如果终端用户在 30s 内没有得到系统的响应,则会变得不耐烦,改变分析主线索,影响分析的质量。

(2) 可分析性(analysis): OLAP 系统应能处理与应用有关的任何逻辑分析和统计分析。尽管系统需要一些事先的编程,但并不意味着系统事先已将所有的应用都定义好了。

(3) 多维性(multidimensional): 多维性是 OLAP 的特点。系统必须提供对数据分析的多维视图和分析,包括对层次维和多重层次维的完全支持。

(4) 信息性(information): 不论数据量有多大,也不管数据存储在何处,OLAP 系统都应能及时获得信息,并且管理大容量的信息。

用于实现 OLAP 的技术主要包括网络环境上客户机/服务器体系结构、时间序列分析、面向对象、并行处理、数据存储优化等。

3.1.2 OLAP 准则

1985 年以来,关系数据库需求始终受到 E. F. Codd 提出的 12 条规则的影响。1993 年, E. F. Codd 在 *Providing OLAP to User Analysts* 一书中又提出了有关 OLAP 的十二条准则,用来评价分析处理工具,这也是他继关系数据库和分布式数据库提出的两个“十二条准则”后提出的第三个“十二条准则”。由于这些规则最初是对客户研究的结果,所以业界对这个十二条准则褒贬不一。但其主要方面,如多维数据分析、客户/服务器结构、多用户支持及一致的报表性能等方面还是得到了大多数人的认可。E. F. Codd 在文中系统阐述了有关 OLAP 产品及其所依赖的数据分析模型的一系列概念及衡量标准,这对 OLAP 产品的辨别

及后来的发展方向的确立都产生了重要的作用。如今,这十二条规则也成为大家定义 OLAP 的主要依据,被认为是 OLAP 产品应该具备的特征。如今 OLAP 的概念已经在商业数据库领域得以广泛使用,Codd 提出的 OLAP 准则如下。

1. 多维概念视图

从用户分析员的角度来看,用户通常从多维角度来看待企业,企业决策分析的目的不同,决定了分析和衡量企业的数据总是从不同的角度来进行的,所以企业数据空间本身就是多维的。因此 OLAP 的概念模型也应是多维的。用户可以简单、直接地操作这些多维数据模型。例如,用户可以对多维数据模型进行切片、切块、改变坐标或旋转模式中的联合(概括和聚集)数据路径。

2. 透明性

透明性原则包括两层含义:首先,OLAP 在体系结构中的位置对用户是透明的。OLAP 应处于一个真正的开放系统结构中,它可使分析工具嵌入用户所需的任何位置,而不会对分析工具的使用产生副作用,同时必须保证 OLAP 工具的嵌入不会引入和增加任何复杂性。其次,OLAP 的数据源对用户也是透明的。用户只需使用熟悉的查询工具进行查询,而不必关心 OLAP 工具获取的数据是来自于同质还是异质的数据源。

3. 可访问性

OLAP 系统不仅能进行开放的存取,而且还提供高效的存取策略。OLAP 用户分析员不仅能在公共概念视图的基础上对关系数据库中的数据进行分析,而且还可以在公共分析模型的基础上对关系数据库、数据仓库的数据进行分析。要实现这些功能,就要求 OLAP 能将自己的概念视图映射到异质的数据存储上,并可访问数据,还能进行所需的转换以便给出单一的、连贯的、一致的用户视图。另外必须说明的一点就是,物理数据来源于何种系统,这对用户来说应是透明的,进行处理的是 OLAP 工具而不是用户分析员。这是提供 OLAP 工具透明性准则的基础之一。

OLAP 系统应该提供高效的存储策略,使系统只存取与指定分析有关的数据,避免多余的数据存取。

4. 一致稳定的报表性能

报表操作不应随维数增加而削弱,即当数据维数和数据的综合层次增加时,提供给最终分析员的报表能力和响应速度不应该有明显的降低,这对维护 OLAP 产品的简易性至关重要。即便是企业模型改变,关键数据的计算方法也无需更改。也就是说,OLAP 系统的数据模型对企业模型应该具有“鲁棒”性。只有做到这一点,OLAP 工具提供的数据报表和所做的预测分析的结果才是可信的。

5. 客户/服务器体系结构

OLAP 是建立在客户/服务器体系结构之上的。这要求它的多维数据库服务器能够被

不同的应用和工具所访问,服务器端以最小的代价完成同多种服务器之间的挂接任务,智能化服务器必须具有在不同的逻辑的和物理的数据库间映射并组合数据的能力,还应构造通用的、概念的、逻辑的和物理的模式,从而保证透明性和建立统一的公共概念模式、逻辑模式和物理模式。客户端负责应用逻辑及用户界面。

6. 维的等同性

每一数据维在其结构和操作功能上必须等价。可能存在适用于所有维的逻辑结构,提供给某一维的任何功能也应提供给其他维,即系统可以将附加的操作能力授给所选维,但必须保证该操作能力可以授给任意的其他维,即要求维上的操作是公共的。该准则实际上是对维的基本结构和维上的操作的要求。

7. 动态的稀疏矩阵处理

OLAP 服务器的物理结构应完全适用于特定的分析模式,创建和加载此种模式是为了提供优化的稀疏矩阵处理。当存在稀疏矩阵时,OLAP 服务器应能推知数据是如何分布的,以及怎样存储才更有效。

该准则包括两层含义:第一,对任意给定的稀疏矩阵,存在一个最优的物理视图,该视图能提供最大的内存效率和矩阵处理能力,稀疏度是数据分布的一个特征,不能适应数据集的数据分布,将会导致快速、高效操作的失败。第二,OLAP 工具的基本物理数据单元可配置给可能出现的维的子集。同时,还要提供动态可变的访问方法并包含多种存取机制,例如:直接计算地址、B 树索引、导出算法、哈希算法或这些技术的最佳组合。访问速度不会因数据维的多少、数据集的大小而变化。

如果分析要求较为单一和固定,那么确实有可能针对它建立起一个最优的、静态的、具有固定维数的物理模式。但实际上,分析需求的特点就是具有不确定性,所以建立静态模式是不现实的,因此 OLAP 工具必须使得模型的物理模式充分适应指定的维数,尤其是特定模型的数据分布。

8. 多用户支持能力

当多个用户在同一分析模式上并行工作,或是在同一企业数据上建立不同的分析模型时,OLAP 工具应提供并发访问、数据完整性及安全性等功能。

实际上,OLAP 工具必须支持多用户也是为了适合数据分析工作的特点。应该鼓励以工作组的形式来使用 OLAP 工具,这样多个用户就可以交换各自的想法和分析结果。

9. 非限定的跨维操作

在多维数据分析中,所有维的生成和处理都是平等的。OLAP 工具应能处理维间相关计算。如果计算时需要按语言定义各种规则,此种语言应允许计算和数据操作跨越任意数目的数据维,而不必限制数据单元间的任何关系,也不必考虑每一单元包含的通用数据属性数目。

10. 直观的数据操作

OLAP 操作要求直观易懂。如果要重定向联系路径,或在维或行间进行细剖操作,都应该通过直观的操作分析模型来完成,而不需要使用菜单,也不需要跨越用户界面进行多次操作,即综合路径重定位、向上综合、向下钻取和其他操作都可以通过直观、方便的点、拉操作来完成。

在分析模型中定义的维应包含用户分析所需的所有信息,从而可以进行任意继承操作。

11. 灵活的报表生成

使用 OLAP 服务器及其工具,用户可以按任何想要的方式来操作、分析、综合和查看数据,这些方式包括将行、列及单元按需要依次排放。报表机制也应提供此种灵活性,报表必须能从各种可能的方面显示出从数据模型中综合出的数据和信息,充分反映数据分析模型的多维特征,并可按用户需要的方式来显示它。

12. 不受限制的维和聚集层次

OLAP 服务器应能在一通用分析模型中协调至少 15 个维。每一通用维应能允许有任意个用户定义的聚集,而且用户分析员可以在任意给定的综合路径上建立任意多个聚集层次(见 3.4.4 节数据立方体)。

3.1.3 OLAP 的基本概念

OLAP 是针对特定问题的联机数据访问和分析。通过对信息进行快速、稳定一致和交互性的存取,允许管理决策人员对数据进行深入观察。为了对 OLAP 技术有更深入的了解,这里主要介绍在 OLAP 中常用的一些基本概念。

1. 变量

变量是数据的实际意义,即描述数据“是什么”。例如,数据 100 本身并没有意义或者说意义未定,它可能是一个学校的学生人数,也可能是某产品的单价,还可能是某商品的销售量,等等。一般情况下,变量总是一个数值度量指标,例如,“人数”、“单价”、“销售量”等都是变量,而 100 则是变量的一个值。

2. 维

维是人们观察数据的特定角度。例如,企业常常关心产品销售数据随着时间推移而产生的变化情况,这时是从时间的角度来观察产品的销售,所以时间是一个维(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,这时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)。其他还有产品维、顾客维等。

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,

通常称这多个描述方面为维的层次。一个维往往具有多个层次。例如,描述时间维时,可以从日期、月份、季度、年等不同层次来描述,那么日期、月份、季度、年等就是时间维的层次;同样,城市、地区、国家等构成了地理维的层次。

4. 维成员

维的一个取值称为该维的一个维成员。如果一个维是多层次的,那么该维的维成员是由各个不同维层次的取值组合而成的。例如,考虑时间维具有日期、月份、年这三个层次,分别在日期、月份、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。对应一个数据项来说,维成员是该数据项在某维中位置的描述。例如对一个销售数据来说,时间维的维成员“某年某月某日”就表示该销售数据是“某年某月某日”的销售数据,“某年某月某日”是该销售数据在时间维上位置的描述。

5. 多维数组

一个多维数组可以表示为:(维 1,维 2,...,维 n ,变量)。例如,若日用品销售数据是按时间、地区和销售渠道组织起来的三维立方体,加上变量“销售额”,就组成了一个多维数组(地区,时间,销售渠道,销售额),如果在此基础上再扩展一个产品维,就得到一个四维的结构,其多维数组为(产品,地区,时间,销售渠道,销售额)。

6. 数据单元(单元格)

多维数组的取值称为数据单元。当多维数组的各个维都选中一个维成员时,这些维成员的组合就唯一确定了一个变量的值。那么数据单元就可以表示为:(维 1 维成员,维 2 维成员,...,维 n 维成员,变量的值)。例如,在产品、地区、时间和销售渠道上各取维成员“牙膏”、“上海”、“2004 年 12 月”和“批发”,就唯一确定了变量“销售额”的一个值(假设为 100 000),则该数据单元可表示为(牙膏,上海,2004 年 12 月,批发,100 000)。

3.2 OLAP 的数据模型

建立 OLAP 的基础是多维数据模型,多维数据模型的存储可以有多种不同的形式。MOLAP 和 ROLAP 是 OLAP 的两种主要形式,其中 MOLAP(Multi-dimension OLAP)是基于多维数据库的 OLAP,简称为多维 OLAP;ROLAP(Relation OLAP)是基于关系数据库的 OLAP,简称关系 OLAP。还有几种 OLAP,如 WOLAP(Web OLAP)代表网络 OLAP,HOLAP(Hybrid OLAP)代表混合 OLAP。

3.2.1 MOLAP 数据模型

MOLAP 数据模型是基于多维数据库的 OLAP,多维数据库(Multi Dimensional DataBase,MDDb)是以多维方式组织数据,即以维作为坐标系,采用类似于数组的形式存储

数据。多维数据库中的元素具有相同类型的数值,如销售量。例如,二维 MDDB(数组,即矩阵)的数据组织如表 3.1 所示。它代表不同产品(衣服、鞋、帽)在不同地区(北京、上海、广州)的销售量情况。

表 3.1 MDDB(二维)数据组织

	北京	上海	广州
衣服	600	700	500
鞋	800	900	700
帽子	100	200	80

在查询中除查询一般的“衣服在广州的销售量”外,有时查询像“衣服的总销售量”等类问题,它涉及多个数据项求和,如果采取临时进行累加计算,会使查询效率大大降低。为此,需要增加汇总数据项。在多维数据库中只需要按行或列进行求和,增加“总和”的维成员即可,如表 3.2 所示。

表 3.2 多维数据库中含综合数据的数据组织

	北京	上海	广州	总和
衣服	600	700	500	1800
鞋	800	900	700	2400
帽子	100	200	80	380
总和	1500	1800	1280	4580

MDDB 的数据组织形式不同于关系数据库的组织形式,关系数据库是以“属性一元组(记录)”形式组织数据。对表 3.1 中的数据按关系数据库组织,数据如表 3.3 所示。

表 3.3 关系数据库 RDBMS 数据组织

产品名	地区	销售量	产品名	地区	销售量
衣服	北京	600	鞋	广州	700
衣服	上海	700	帽子	北京	100
衣服	广州	500	帽子	上海	200
鞋	北京	800	帽子	广州	80
鞋	上海	900			

可见,多维数据库 MDDB 比关系数据库表达更清晰且占用的存储少。在关系数据库中增加综合数据项,如表 3.4 所示。这些综合数据项一般在建立数据库的同时计算出来。这样在查询时,不必临时进行计算,提高了查询效率。对于多维数据库的综合数据项明显比关系数据库的综合项更有效果。

表 3.4 关系数据库中综合数据的数据组织

产品名	地区	销售量	产品名	地区	销售量
衣服	北京	600	鞋	广州	700
衣服	上海	700	鞋	总和	2400
衣服	广州	500	帽子	北京	100
衣服	总和	1800	帽子	上海	200
鞋	北京	800	帽子	广州	80
鞋	上海	900	帽子	总和	380

3.2.2 ROLAP 数据模型

ROLAP 是基于关系数据库的 OLAP,如表 3.3 所示。它是一个平面结构,用关系数据库表示多维数据时,采用星型模型,即用两类表,一类是事实表,存储事实的实际值,如销售量;另一类是维表,对每一个维来说,至少有一个表来存储该维的描述信息,如产品的名称、分类等。星型模型完全用二维关系表示了数据的多维观念。

通过关系数据库实现多维查询时,通过维表的主码对事实表和每一个维表做连接操作,一次查询就可以得到数据的具体值以及对数据的多维描述(即对应的各维上的维成员)。但是,因为对每个维都需要进行一次连接操作,所以系统的性能就成了 ROLAP 实现的最大的一个问题,特别是当维数增加和事实表增大时,必须采用有效的查询优化技术(特别是表连接策略),利用各种索引技术来提高系统的性能。

当存在多层次的复杂维时,需要采用“雪花模型”,用多张表来描述一个复杂维。对于存在综合数据时,需要建立汇总事实表,采用“星网模型”来描述。

3.2.3 MOLAP 与 ROLAP 的比较

MOLAP 通过多维数据库引擎从关系数据库 DB 和数据仓库 DW 中提取数据,将各种数据组织成多维数据库,存放到 MDDB 中,而且将自动建立索引并进行预综合(见 3.4.4 节)来提高查询存取性能,如图 3.1 所示。

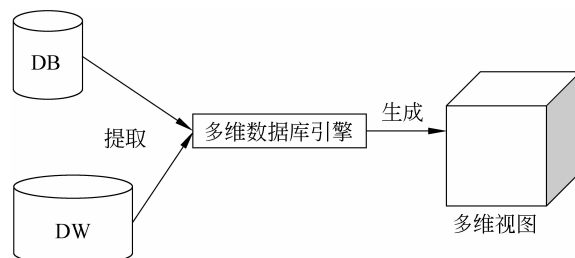


图 3.1 MOLAP 结构

ROLAP 从关系数据库 DB 和数据仓库 DW 中提取数据,按关系 OLAP(ROLAP)的数据组织存放在关系数据库服务器(RDBMS 服务器)中。最终用户的多维分析请求,通过

ROLAP 服务器的多维分析(OLAP)引擎动态翻译成 SQL 请求,将查询结果经多维处理(将关系表达式转换成多维视图)返回用户,如图 3.2 所示。

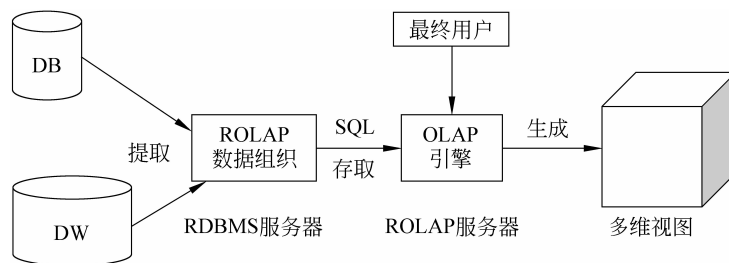


图 3.2 ROLAP 结构

虽然这两种技术都满足了 OLAP 数据处理的一般过程,即数据装入、汇总、建索引和提供使用,但 MOLAP 要比 ROLAP 简明一些,MOLAP 的索引及数据综合可以自动进行。然而 ROLAP 的实现较为复杂,但灵活性较好,用户可以动态实现统计或计算方式。

下面详细深入分析 MOLAP 与 ROLAP。

1. 数据存取速度

ROLAP 的多维数据是以星型模型等关系数据库(平面形式)存储,并不直接体现“超立方体”形式。在接收客户 OLAP 请求时,ROLAP 服务器需要将 SQL 语句转化为多维存储语句,并利用连接运算临时“拼合”出多维数据立方体。因此,ROLAP 的响应时间较长。

目前,关系型数据库已经对 OLAP 做了很多优化,包括并行存储、并行查询、并行数据管理、基于成本的查询优化、位图索引、SQL 的 OLAP 扩展等,大大提高了 ROLAP 的速度。

MOLAP 是专为 OLAP 所设计的,能够自动地建立索引,并且有良好的预计算能力,能够使用多维查询语句访问数据立方体,因此 MOLAP 在数据存储速度上性能好,响应速度快。

2. 数据存储的容量

ROLAP 使用的传统关系数据库的存储方法,在存储容量上基本没有限制。但是,需要指出的是,在 ROLAP 中为了提高分析响应速度,常常构造大量的中间表(如综合表),这些中间表带来了大量的冗余数据。

MOLAP 通常采用多平面叠加成立体的方式存放数据,(这样访问速度快),由于受操作系统平台中文件大小的限制,当数据量超过操作系统最大文件长度时,需要进行数据分割。随着数量的增大,多维数据库进行的预运算结果将占用巨量的空间,此时可能会导致“数据爆炸”的现象。因此,多维数据库的数据量级难以达到太大的字节级。

3. 多维计算的能力

MOLAP 能够支持高性能的决策支持计算,包括复杂的跨维计算、行级的计算,而在 ROLAP 中,SQL 无法完成部分计算,并且 ROLAP 无法完成多行的计算和维之间的计算。

最近发展起来的多维数据分析语言 MDX 能更有效地进行多维数据分析(见 3.4.5 节)。

4. 维度变化的适应性

MOLAP 需要在建立多维数据库前确定各个维度以及维度的层次关系。在多维数据库建立之后,如果要增加新的维度,则多维数据库通常需要重新建立。新增维度数据会剧烈增加。而 ROLAP 增加一个维度,只是增加一张维表并修改事实表,系统中其他维表不需要修改,因此 ROLAP 对于维表的变更有很好的适应性。

5. 数据变化的适应性

由于 MOLAP 通过预综合处理来提高速度,当数据频繁地变化时,MOLAP 需要进行大量的重新计算,甚至重新建立索引乃至重构多维数据库。在 ROLAP 中,预综合处理通常由设计者根据需求制定,因此灵活性较好,对于数据变化的适应性强。

6. 软硬件平台的适应性

关系数据库已经在众多的软硬件平台上成功地运行,即 ROLAP 对软硬件平台的适应性很好,而 MOLAP 相对较差。

7. 元数据管理

元数据是 OLAP 和数据仓库的核心数据,OLAP 的元数据包括层次关系、计算转化信息、报表中的数据项描述、安全存取控制、数据更新、数据源和预计算综合表等,目前在元数据的管理上,MOLAP 和 ROLAP 都没有成形的标准,MOLAP 产品将元数据作为其内在数据,而 ROLAP 产品将元数据作为应用开发的一部分,由设计者来定义和处理。

MOLAP 和 ROLAP 在技术上各有优缺点。MOLAP 以多维数据库为核心,在数据存储和综合上有明显的优势,但它不适应太大的数据存储,特别是对有大量稀疏数据的存储将会浪费大量的存储空间。ROLAP 以 RDBMS 为基础,利用成熟的技术为用户的使用和管理带来方便。

MOLAP 和 ROLAP 在数据存储、技术和特性方面的比较如表 3.5 所示。

表 3.5 MOLAP 和 ROLAP 的比较

	数 据 存 储	技 术	特 征
MOLAP	详细数据用关系表存储在数据仓库中; 各种汇总数据保存在多维数据库中; 从数据仓库中询问详细数据,从多维数据库中询问汇总数据	由 MOLAP 引擎创建; 预先建立数据立方体; 多维视图存储在陈列中,而不是表格中; 可以高速检索矩阵数据; 利用稀疏矩阵技术来管理汇总的稀疏数据	询问响应速度快; 能轻松适应多维分析; 有广泛的下钻和多层次/多视角的查询能力
ROLAP	全部数据以关系表存储在数据仓库中; 可获得细节的和综合汇总的数据; 有非常大的数据容量; 从数据仓库中询问所有的数据	使用复杂 SQL 从数据仓库中获取数据; ROLAP 引擎在分析中创建数据立方体; 表示层能够表示多维的视图	在复杂分析功能上有局限性,需要采用优化的 OLAP; 向下钻取较容易,但是跨维向下钻取比较困难