联机分析处理

1 OLAP 概念

1.1 OLAP 定义

OLAP 理事会给出的定义

联机分析处理是一种软件技术，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。

OLAP 简单定义

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 4 个特征：

1) 快速性：用户对 OLAP 的快速反应能力有很高的要求。系统应在 5 秒内对用户的大部分分析要求做出反应。

2) 可分析性：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对所有的应用都定义好了。

3) 多维性：OLAP 的关键属性，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：OLAP 系统不论数据量有多大，也不管数据存储在何处，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 C/S 体系结构、时间序列分析、面向对象、并行处理、数据存储优化以及多线索技术等。

1.2 OLAP 准则

1993 年，E.F. Codd 在《Providing OLAP to User Analysts》中提出了有关 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“十二条条准则”。

1) 多维概念视图

从用户分析员的角度来看，用户通常按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，所以企业数据空间本身就是多维的。因此 OLAP 的概念模型也就是多维的。

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 应处理一个真正开放系统结构中，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，而且还能提供高效的存取策略。

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。

5) 客户/服务器体系结构

OLAP 是建立在客户/服务器体系结构上的，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，服务器端智能地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。

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型，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析的，以及怎样存储才更有效。

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。

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，此种语言应允许吉“数据的计算和数据的操作”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，惑乱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。

11) 灵活的报表生成

用户可以用 OLAP 服务器及其工具，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、列和单元按需要依次排序。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”应能允许有任意多个用户定义的聚集，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。

1.3 OLAP 基本概念

变量：数据的实际意义，描述数据“是什么”。如数据 100，可以把它定义为“人数”，

一般情况下，变量是一个数值量指标。

维：人们观察数据的特定角度。时间维，地理维，产品习以为常，顾客维等。

维的层次： 某个特定角度还可以存在细节程序不同的多个描述方面， 称这多个描述方面为维的层次。一个“ 维 ”往往具有多个层次，如时间维，可以从日期、月份、季度、年等不同层次来描述。

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 ”由各个不同维层次的取值组合而成。

多维数组：一个多维数组可表示为： （维 1，维 2，…，维 n，变量）。如，日用品销售数据是按时间、地点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） 。

数据单元（单元格） ：多维数组的取值称为数据单元格。当多维数组的各个维都选中一个维成员，就确定了一个变量的值。

2 OLAP 的数据模型

2.1 MOLAP 数据模型

MOLAP 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，采用类似于数组形式存储数据。

MDDB （二维）数据组织如下表

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

带有综合数据的数据组织

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

多维数据库组织形式不同于关系数据库组织形式， 关系数据库是以“ 属性 -元组 ”形式记录数据。如图：

产品名	地区	销售量
衣服	北京	600
衣服	上海	700
衣服	广州	500
鞋	北京	800
鞋	上海	900

鞋	广州	700
帽子	北京	100
帽子	上海	200
帽子	广州	80

关系数据库带“综合项”的数据组织形式，如图：

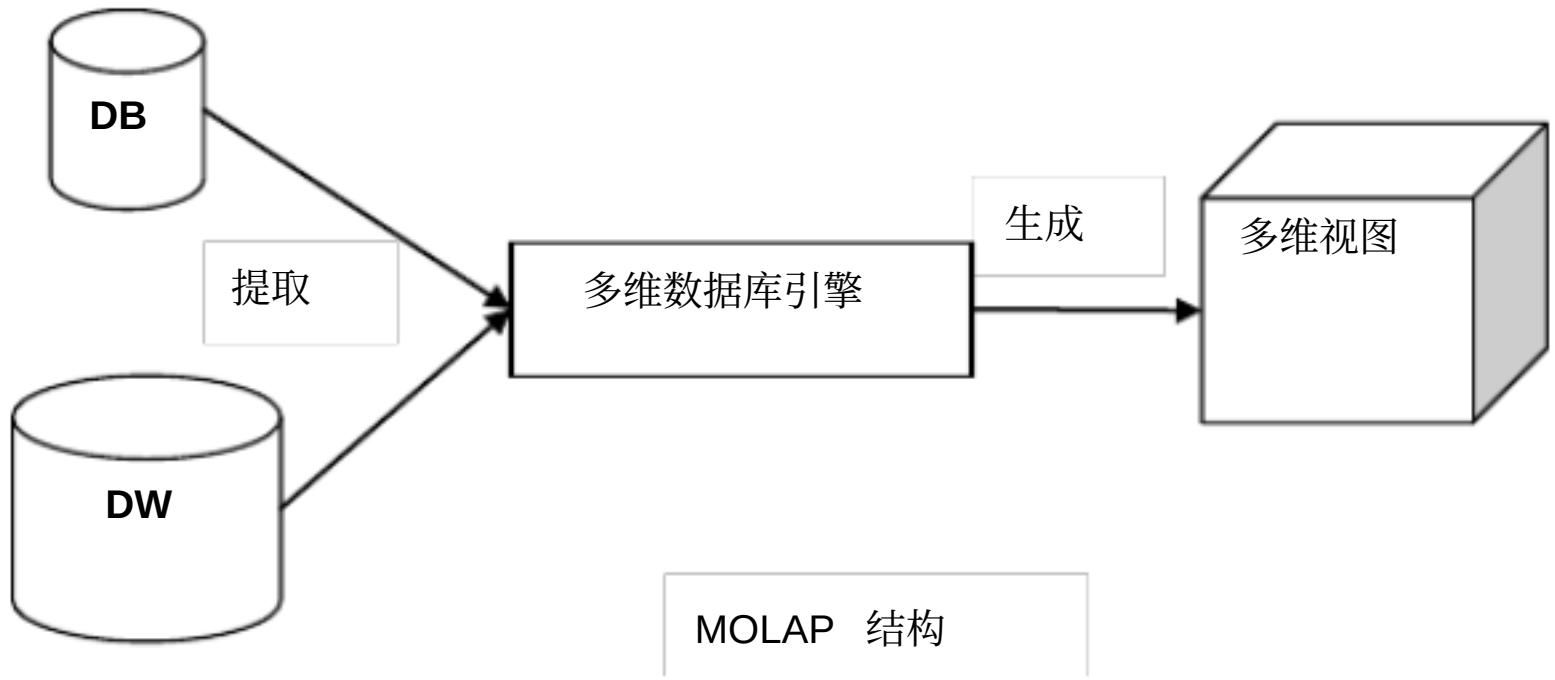
产品名	地区	销售量
衣服	北京	600
衣服	上海	700
衣服	广州	500
衣服	总和	1800
鞋	北京	800
鞋	上海	900
鞋	广州	700
鞋	总和	2400
帽子	北京	100
帽子	上海	200
帽子	广州	80
帽子	总和	380

2.2 ROLAP 数据模型

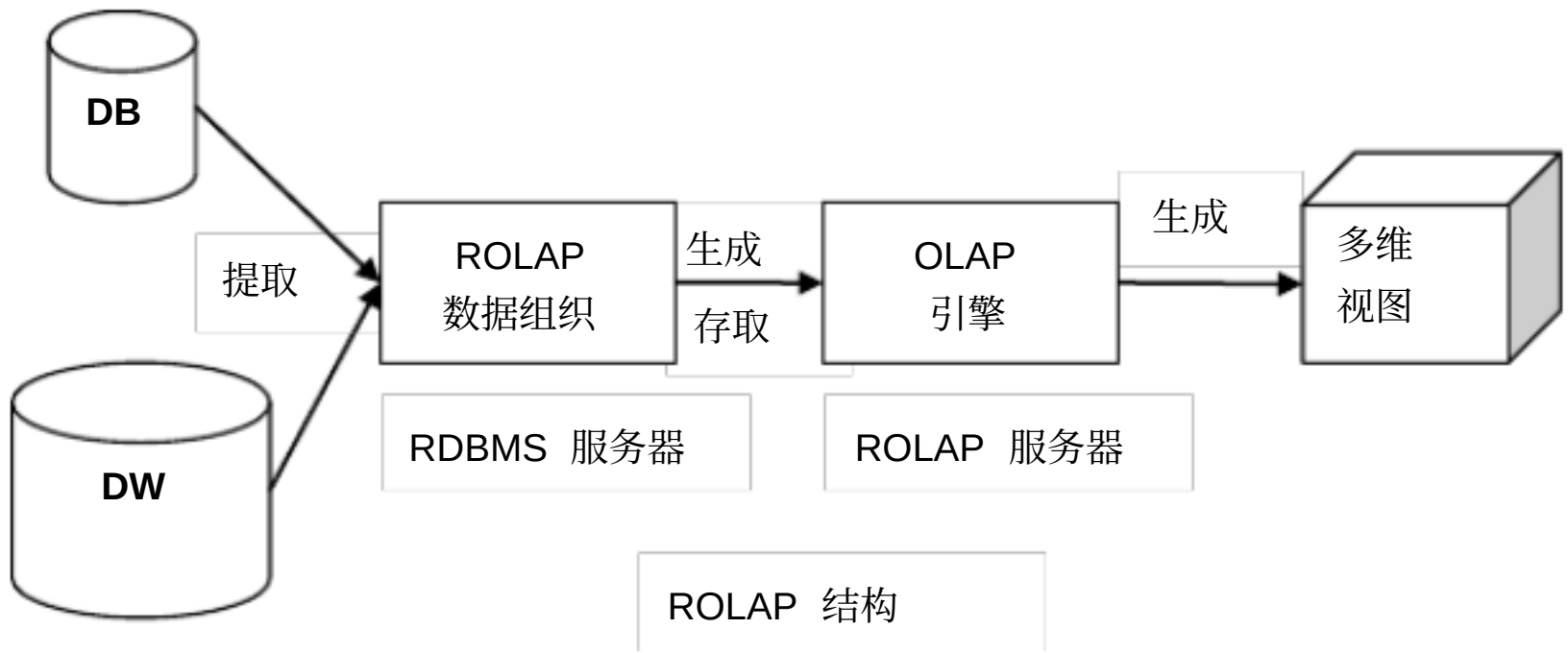
ROLAP 是基于关系数据库的 OLAP，见上节的关系数据库的数据组织形式。它是一个平面结构，用关系数据库表示多维数据时，采用星型模型，即用两类表，一类是事实表，存储事实的实际值，如销售量；另一类是维表，对每一个维来说，至少有一个表来存储该“维”的描述信息，如产品的名称、分类等。由于关系数据库实现多维查询时，应进行查询优化技术（特别是表连接策略），利用各种索引技术来提高系统的性能。ROLAP 常用星型模型或雪花模型来创建数据逻辑模型。

2.3 MOLAP 与 ROLAP 的比较

MOLAP 与 ROLAP 的结构差别，如下图：



MOLAP 结构



ROLAP 结构

两者对比如下表

特性	ROLAP	MOLAP
数据存取速度	平面型式存储，慢	数据立方体，快
数据存储的容量	容量大（冗余多）	容量小（冗余少）
多维计算的能力	无法多行和“维”之间计算	高性能计算，较强
维度变化的适应性	较强	较差，（多维结构特点）
数据变化的适应性	较强	很差，（大量重新计算）
软硬件平台的适应性	较强	较高（多维特殊性）
元数据管理	应用开发一部分， 要定义处理	作为其内部数据

两者在数据存储、技术和特性的比较：

项目	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；各汇总数据保存在多维	由 MOLAP 引擎创建；预先建立数据立方体；多维视图存储	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

	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	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	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

2.4 HOLAP 数据模型

HOLAP （hybrid OLAP ），即混合 OLAP 介于 MOLAP 和 ROLAP 之间。

3 多维数据的显示

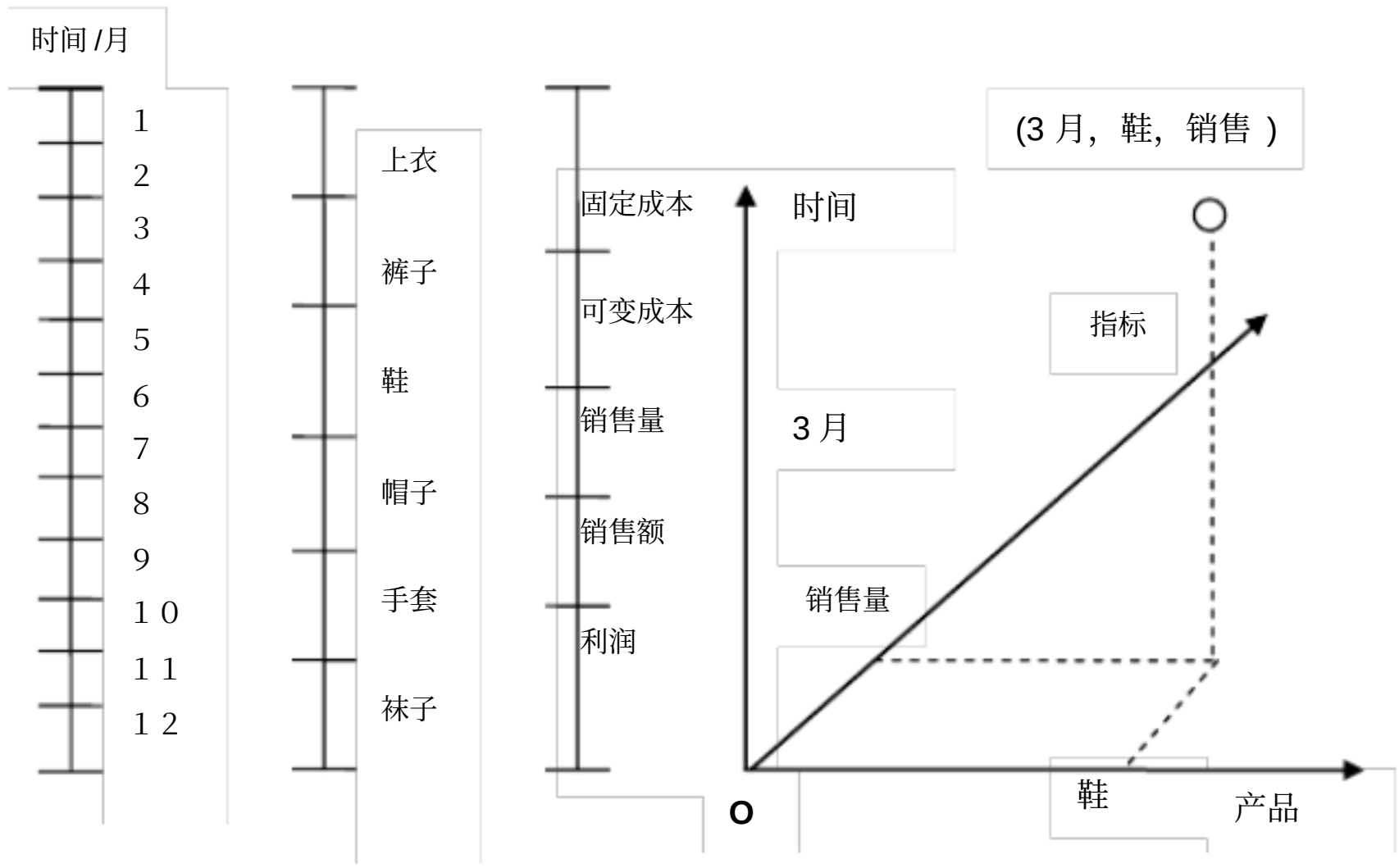
3.1 多维数据的显示方法

多维数据都要在平面上显示出来，所以在显示“ 三维 ”以上数据时，常要固定一些维成员，重点显示两维的数据，如下表形式：

北京地区	1 月	2 月	3 月
衣服	100	200	300
鞋	150	300	350

3.2 多维类型结构

E.Thomsen 引入多维类型结构（ MTS ），有些专家称为多维域结构（ MDS ）。表示方法是：每一个维度用一条线段来表示。维度中的每一个成员都用线段上的一个单位区间来表示。图



3.3 多维数据的分析视图

在平面的屏幕上显示多维数据，是利用行、列和页面 3 个“ 显示组 ”来表示的。对于更多维度的数据显示，需要选择维度及其成员分布在行或列中。在页面上可以选定多个维度，但每个“ 维度 ”只能显示一个成员。在行或者列中一般只选择 2 个维，每个“ 维 ”可以有多个成员。 由于整个屏幕的空间是有限的， 将维度嵌套在行或者列中相对于放在页维度中会占据更多的屏幕空间。一些有经验的规则如下：

将维度尽量放在页中， 除非确定需要同时看到一个维度的多个成员。 让民屏幕上的信息尽量相关。

当维度嵌套在行或者列中时， 考虑到垂直空间比水平空间更有用， 所以将维度嵌套在列中比嵌套在行中要好。一个经典的显示主法就是在行上有一个维度，而在列上嵌套 1 到 3 个维度，而其他的维度则放在页中。

在决定 数据的屏幕显示方式之前，应该首先弄清楚需要查找和分析比较的内容。如，如果需要比较某一个产品和某类客户在商品和时间上的实际成本情况， 可以将产品和客户放在页面维度中，而在屏幕上则可以按商店和时间来显示实际成本。

4 维数据的显示：

商店 3 （页面）	上衣		裤子		帽子	
	直接销售	固定成本	直接销售	固定成本	直接销售	固定成本
1 月	450	350	550	450	500	400
2 月	380	280	460	360	400	320
3 月	400	310	480	410	450	400

6 维数据的显示：

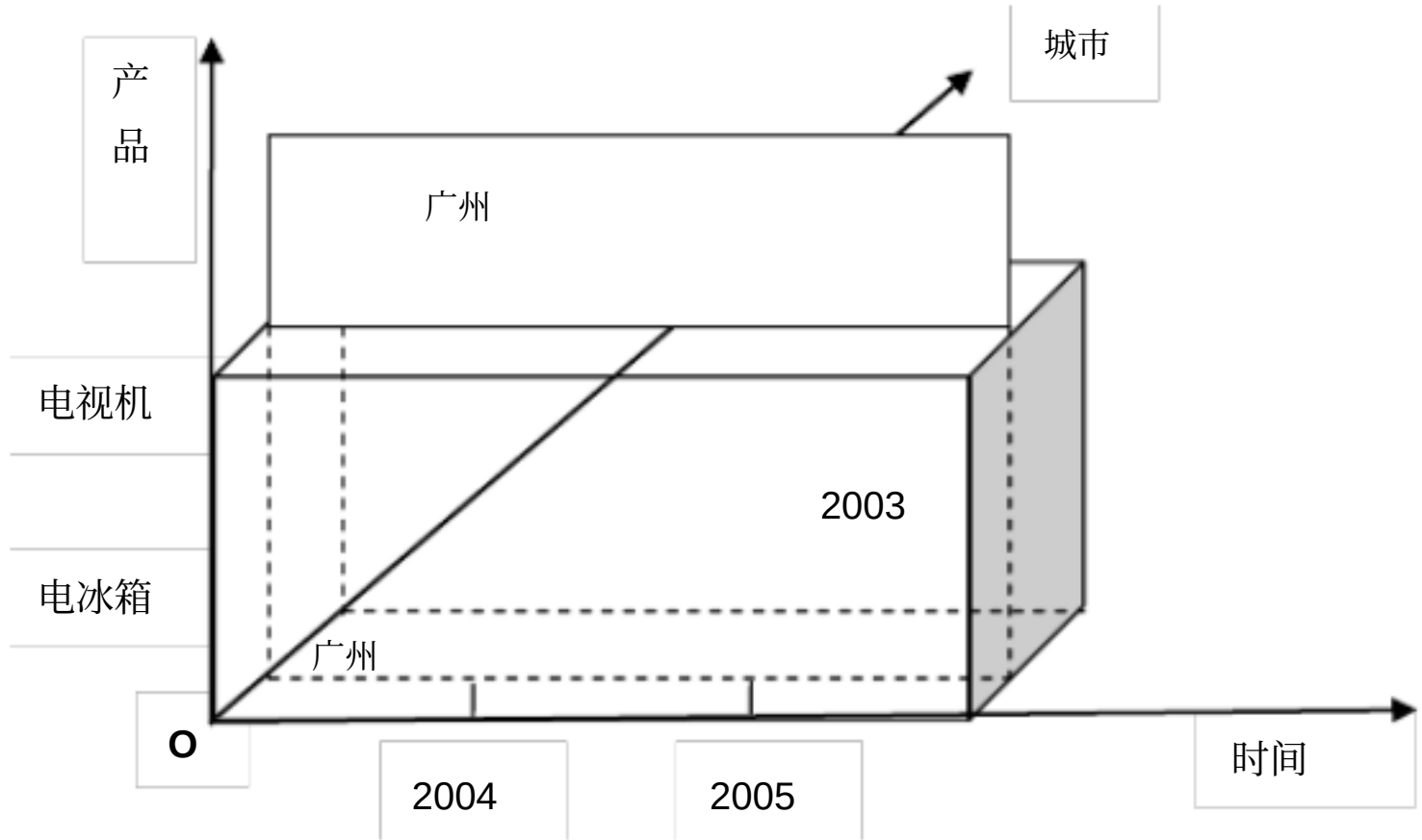
商店 3,老年 （页面）	直接销售		间接销售		总销售	
	实际	计划	实际	计划	实际	计划

1 月	桌子	250	300	125	160	375	450
	台灯	265	320	133	160	400	480
2 月	桌子	333	400	167	200	500	600
	台灯	283	340	142	170	425	510
3 月	桌子	350	420	175	210	525	630
	台灯	250	300	125	150	375	450

4 OLAP 的多维数据分析

4.1 多维数据分析的基本操作

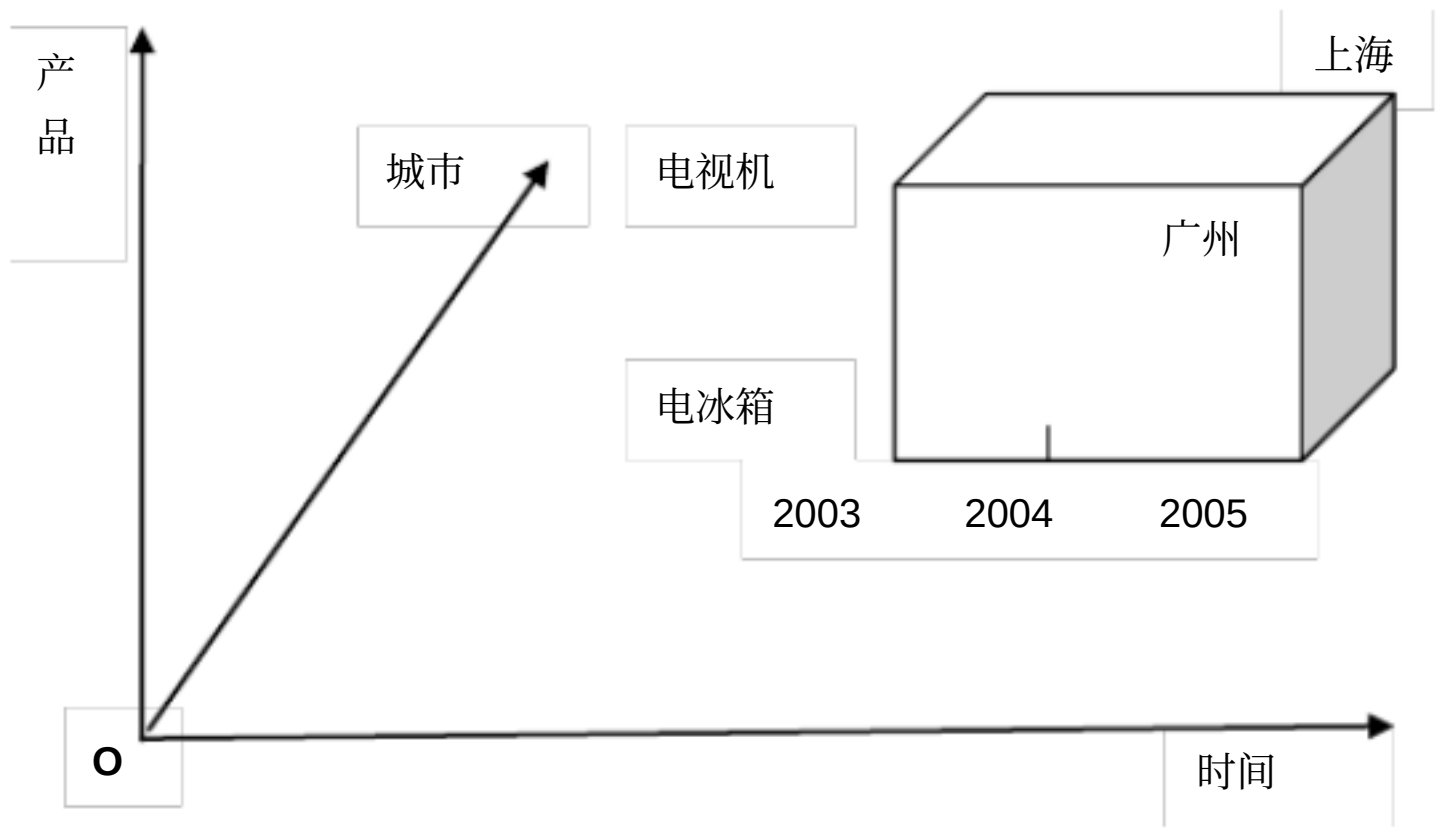
1) 切片 **(slice)**：在某两个“ 维 ”上取一定区间的“ 维成员 ”或全部“ 维成员 ” ，而在其余的“ 维 ”上选定“ 维成员 ”的操作。如下图三维数据立方体（地区，时间，产品，销售额）。如果在地区“ 维 ”上选定一个“ 维成员 ” （设为上海），就得到了在地区上的一个切片（在于时间和产品的切片） ；这样的切片数目取决于每个“ 维 ”上“ 维成员 ”的个数。



2) 切块 **(dice)**：有两种

在多维数组的某一个“ 维 ”上选定某一区间的“ 维成员 ”的操作。也即是多个切片叠合起来。如果将时间“ 维 ”上的取值设定为一个区间（如 2001-2005 年），就得到一个数据切块，可以看成由 2001 年至 2005 年 5 个切片叠合而成。

选定多维数组给的一个三维子集的操作。 在多维数组 （维 1， 维 2， …维 n，变量）中选定 3 个“ 维 ” ，维 i，维 j，维 k，在这 3 个“ 维 ”上分别取一个区间，或任意“ 维成员 ”，而其他“ 维 ”都取定一个“ 维成员 ” 。



3) 钻取 **(drill)** ：有向下钻取（ drill down）和向上钻取 (drill up)操作。向下钻取是使用户在多层数据中能通过导航信息而获得更多的细节性数据, 而向上钻取获得概括性的数据。钻取深度与“ 维 ”所划分的层次相对应。如 2005 年各部门销售收，如下表

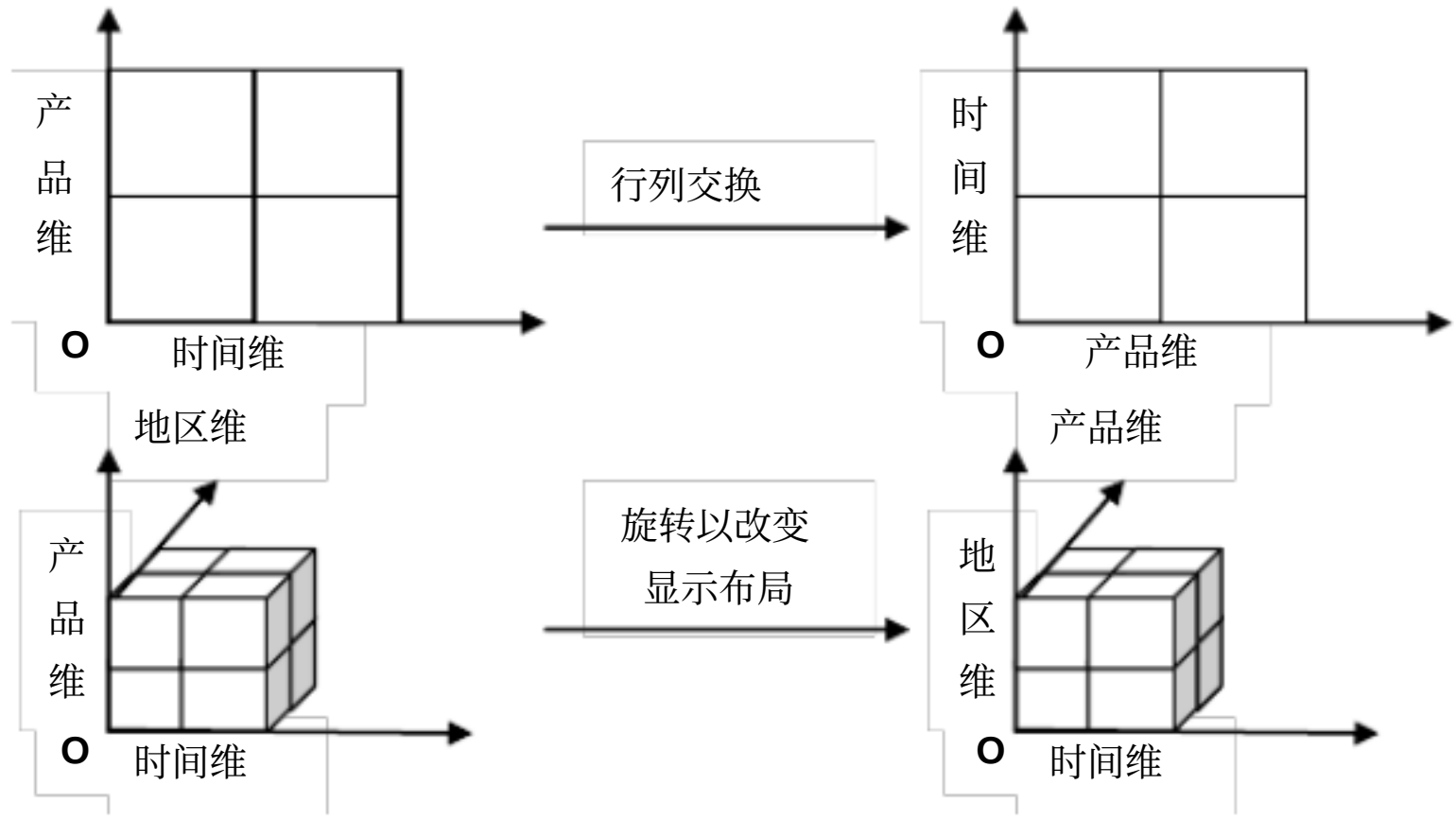
部门	销售
部门 1	900
部门 2	650
部门 3	800

在“ 时间维 ”进行下钻操作

项目	2005 年			
部门	1 季度	2 季度	3 季度	4 季度
部门 1	200	200	350	150
部门 2	250	50	150	150
部门 3	200	150	180	270

向上钻取，正好是上面相反的操作。

4) 旋转 **(pivot)**：通过旋转可以得到不同视角的数据。 旋转操作相当于平面数据将坐标轴旋转。如可能包含了交换行和列，或是把某一个行“ 维 ”移到列“ 维 ”中去，或是把页面显示中的一个维和页面外的“ 维 ”进行交换（令其成为新的行或列中的一个） 。如下图



4.2 广义 OLAP 功能

从广义上讲，任何有助于辅助用户理解数据的技术或者操作都可以作为 OLAP 功能，这些有别于基本 OLAP 的功能（切片，切块，钻取，旋转等）称为广义 OLAP 功能。

1) 基本代理操作

“代理”是一些智能性代理，当系统处于某种特殊状态时提醒分析员。

示警报告： 定义一些条件，一量条件满足，系统会提醒分析员去做分析。如每日报告完成或月订货完成等，通知分析员做分析。

时间报告：按日历和时钟提醒分析员。

异常报告： 当超出边界条件时提醒分析员。 如销售情况已超出预定义阈值的上限或下限时提醒分析员。

2) 数据分析模型

以前数据分析主要集中在静态数据值的相互比较上。有了 OLAP 后，可以进行动态数据分析，需要建立企业数据分析模型。 E.F. Codd 将数据分析模型分为 4 类：

绝对模型 （**categorical model**）：属于静态数据分析， 通过比较历史数据值或行为来描述过去发生的事实。 该模型查询比较简单， 综合路径是预先定义好的， 用户交互少。

解释模型（ **exegetical model**）：属于静态数据分析，分析人员利用系统已有的多层次的综合路径层层细化，找出事实发生的原因。

思考模型（ **contemplative model**）：它属于动态数据分析，旨在说明在一维或多维上引入一组具体变量或参数后将会发生什么。 分析人员在引入确定的变量或公式关系时，必须创建大量的综合路径。

公式模型（ **formulaic model**）：它的动态数据分析能力更高，该模型表示在多个维上，需要引入哪些变量或参数，以及引入后所产生的结果。

以一个实例来进行说明：

一家百货公司在立了自己的数据仓库之后，希望构造一个 OLAP 系统辅助决策。决

策者最关心的一个问题是如何最大限度地扩大商品的销售量，　　因而希望能尽可能找出与销售量相关的因素，　从而采取相应的促销手段。　但是能获得多大的帮助取决于采用何种分析模型。

绝对模型只能对历史数据进行比较，并且利用回归分析等一些分析方法得出趋势信息。能回答诸如 “ 某种商品今年的销售情况与以往相比有怎么的变化？今后的趋势怎样？ 等类问题。

解释模型能够　在当前多维视图的基础上找出事件发生的原因。　如该公司按时间、 地区、 商品及销售渠道建立了多维数据库，　假设今年销售量下降，　那么解释模型应当能找出原因，即销售量下降与时间、地区、商品及销售量四者中的何种因素有关。

思考模型在决策者的参与下，　找出关键变量。 如该公司决策者为了了解某商品的销售量是否与顾客的年龄有关引入了行变量——年龄，即在当前的多维视图上增加了顾客的 “ 年龄维 ”。解释模型就能分析出 “ 年龄 ” 的引入是否必要，即商品销售与顾客年龄有关或无关。

公式模型自动完成上述变量引入工作，　从而最终找出与销量有关的全部因素，　并给出引入后的结果。

3)　商业分析模型

利用数据仓库中的数据时行商业分析需要建立一系列模型，用于提高决策支持能力。常用有：分销渠道的分析模型，客户利润贡献度模型，客户关系（信用）优化模型，风险评估模型。

4.3 多维数据分析实例

假设有一个五维数据模型，商店，方案，部门，时间，销售。下面进行分析：

指定 “ 商店＝ ALL ，方案＝现有 ” 情况的三维表（行为部门，列为时间和销售量）　，如下面表中无括号的数为增长率，有括号表示下降率，下同。

商店＝ ALL ，方案＝现有

项目	2004 年		2005 年		增长率 /%	
	销售量	利润增长率	销售量	利润增长率	销售量	利润增长
服装	234670	27.2	381102	21.5	62.4	(20.0)
家具	62548	33.8	66005	31.1	5.6	(8.0)
汽车	375098	22.4	325402	27.2	(13.2)	21.4
所有其他	202388	21.3	306677	21.7	50.7	1.9

对于汽车部门出现的奇怪现象，销售量下降了　13.2%，而利润却增加了　21.4%，此时对汽车部门进行向下钻取，具体项目（维修、附件、音乐）的销售情况和利润增长情况，见下面表

项目	2004 年		2005 年		增长率 /%	
	销售量	利 润 增 长	销售量	利 润 增 长	销售量	利润增长

		率		率		
汽车	375098	22.4	325402	27.2	（13.2）	21.4
维修	195051	14.2	180786	15.0	（7.3）	5.6
附件	116280	43.9	122545	47.5	5.3	8.2
音乐	63767	8.2	22071	14.2	（63.4）	7.3

切片表：去除一部分列或行不显示，如下表

商店＝ ALL ， 方案＝现有

项目	2005 年
	销售量
服装	381102
家具	66005
汽车	325402
所有其他	306677

旋转表：将“ 方案维 ”加入到“ 销售维 ”中。方案维有 3 种情况：现有、计划、最新预测，这次旋转操作得到 2005 年的第一个表中方案的成员有：现有、计划、差量、差量（%）得到旋转表如下：

商店＝ ALL ， 方案＝现有

项目	2005 年			
	销售量			
	现有	计划	差量	差量（%）
服装	381102	350000	31.1	8.9
家具	66005	69000	（2995）	（4.3）
汽车	325402	300000	25402	8.5
所有其他	306677	350000	（43323）	12.7

5 OLAP 结构与分析工具

5.1 OLAP 结构

1) OLAP 逻辑结构

由 OLAP 视图和数据存储组成， OLAP 视图对于用户来说它是数据仓库或数据集市中数据的多维逻辑表示， 不管数据怎样存储和存储在何处； 数据存储要求选择数据实际存储的方式和实际存储位置， 两种常用的选择是多维数据存储和关系数年据存储。

2) OLAP 物理结构

两种方式：多维数据存储和关系数据存储。

5.2 OLAP 的 Web 结构

使用 Web 结构组织 OLAP 应用，如下图：