第5章　数据仓库型决策支持系统

20世纪90年代中期，国外兴起了3项决策支持新技术，即数据仓库、联机分析处理(OLAP)和数据挖掘(DM)。数据仓库是在数据库的基础上发展起来的，把数据的组织由二维平面结构扩充到多维空间结构，用于决策分析。联机分析处理提出了多维数据分析方法。数据挖掘则是在人工智能机器学习中发展起来的，它是从数据库或数据仓库中发现知识(KDD)的核心。数据仓库、联机分析处理、数据挖掘的结合形成了基于数据仓库的决策支持系统。

5.1　数据仓库基本原理

数据仓库是W. H. Inmon在《建立数据仓库》(*Building the Data Warehouse*)中提出的。数据仓库的提出是以关系数据库，并行处理和分布式技术的飞速发展为基础的信息新技术。

从目前的形势看，数据仓库技术已紧跟Internet而上，成为信息社会中获得企业竞争优势的又一关键技术。

5.1.1　数据仓库的概念

1. 数据仓库的定义

1) W. H. Inmon对数据仓库的定义

数据仓库是面向主题的、集成的、稳定的、不同时间的数据集合，用于支持经营管理中的决策制定过程。

2) SAS软件研究的观点

数据仓库是一种管理技术，旨在通过通畅、合理、全面的信息管理，达到有效的决策支持。

传统数据库用于事务处理，也称为操作型处理，是指对数据库联机进行日常操作，即对一个或一组记录的查询和修改，主要为企业特定的应用服务。用户关心的是响应时间、数据的安全性和完整性。数据仓库用于决策支持，也称为分析型处理，用于决策分析，它是建立决策支持系统的基础。操作型数据与分析型数据的对比如表5.1所示。

表5.1　操作型数据与分析型数据对比表

操作型数据	分析型数据	操作型数据	分析型数据
细节的	综合或提炼的	事务驱动	分析驱动
代表当前的数据	代表过去的数据	面向应用	面向分析
可更新的	不更新	一次操作数据量小	一次操作数据量大
操作需求事先可知道	操作需求事先不知道	支持管理业务	支持决策分析

例如，银行的用户有储蓄，又有贷款，还有信用卡。这些数据存放在不同业务处彼此独立的数据库中。现在，把这 3 个数据库集中起来建立数据仓库，就便利对用户的整体分析，容易决定是否继续对用户贷款或发放信用卡。

2. 数据仓库的特点

数据仓库有如下特点。

1）数据仓库是面向主题的

主题是数据归类的标准，每一个主题基本对应一个宏观的分析领域。例如，保险公司的数据仓库的主题为客户、政策、保险金、索赔等。

基于应用的数据库则完全不同，它的数据只是为处理具体应用而组织在一起的。保险公司按应用来组织数据库为汽车保险、生命保险、健康保险、伤亡保险等。

2）数据仓库是集成的

数据进入数据仓库之前，必须经过加工与集成。对不同的数据来源进行数据结构和编码的统一。统一原始数据中的所有矛盾之处，如字段的同名异义、异名同义、单位不统一、字长不一致等。总之，将原始数据结构做一个从面向应用到面向主题的大转变。

3）数据仓库是稳定的

数据仓库中包括了大量的历史数据。数据经集成进入数据仓库后是极少或根本不更新的。

4）数据仓库是随时间变化的

数据仓库内的数据时限在 5～10 年，故数据的键码包含时间项，标明数据的历史时期，这适合决策支持系统进行时间趋势分析。而数据库只包含当前数据，即存取某一时间的正确的有效的数据。

5）数据仓库中的数据量很大

通常的数据仓库的数据量为 10GB 级，相当于一般数据库 100MB 的 100 倍，大型数据仓库是 1TB(1000GB)级。

数据仓库中数据的比重为索引和综合数据占 2/3，原始数据占 1/3。

6）数据仓库软硬件要求较高

既需要一个巨大的硬件平台，又需要一个并行的数据库系统。

5.1.2 数据仓库结构

数据仓库是在原有关系型数据库基础上发展形成的，但不同于数据库系统的组织结构形式，它从原有的业务数据库中获得的基本数据和综合数据被分成一些不同的层次(levels)。一般数据仓库的结构组成如图 5.1 所示，包括当前基本数据(current detail data)、历史基本数据(older detail data)、轻度综合数据(lightly summarized data)、高度综合数据(highly summarized data)和元数据(meta data)。

当前基本数据是最近时期的业务数据，是数据仓库用户最感兴趣的部分，数据量大。当前基本数据随时间的推移，由数据仓库的时间控制机制转为历史基本数据，一般被转存于介

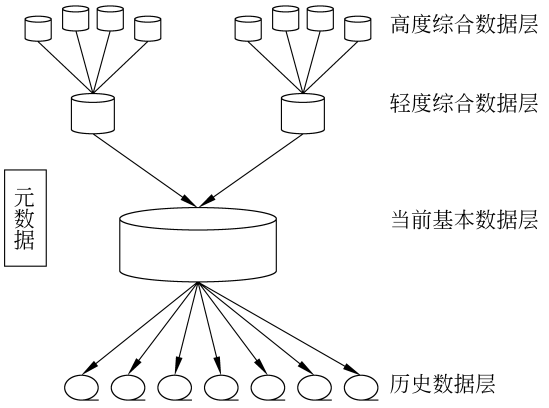

图 5.1　数据仓库结构图

质中，如磁带等。轻度综合数据是从当前基本数据中提取出来的，设计这层数据结构时会遇到“综合处理数据的时间段选取，综合数据包含哪些数据属性（attributes）和内容（contents）”等问题。最高一层是高度综合数据层，这一层的数据十分精练，是一种准决策数据。

整个数据仓库的组织结构是由元数据来组织的，它不包含任何业务数据库中的实际数据信息。元数据在数据仓库中扮演了重要的角色，它被用在以下几种用途。

（1）定位数据仓库的目录作用。

（2）数据从业务环境向数据仓库环境传送时数据仓库的目录内容。

（3）指导从当前基本数据到轻度综合数据，轻度综合数据到高度综合数据的综合方法。元数据至少包括以下一些信息：数据结构（the structure of the data）；用于综合的方法（the algorithms used for summarization）；从业务环境到数据仓库的规划（the mapping from the operation to the data warehouse）。

例如，当前基本数据层存放的是 2015—2016 年销售细节数据，历史数据层存放的 2010—2014 年的销售细节数据，轻度综合数据层存放 2015—2016 年的每周销售数据，高度综合数据层存放 2015—2016 年的每月销售数据。

数据仓库的工作范围和成本常常是巨大的。建造数据仓库需要对所有的用户的任何一次决策需求进行分析，从而使数据仓库的开发成本高、时间长。于是，提供更紧密集成的、拥有完整图形接口并且价格吸引人的工具——数据集市（Data Marts）——就应运而生了。

数据集市是一种更小、更集中的数据仓库，为公司提供分析商业数据的一条廉价途径。

目前，全世界对数据仓库总投资的一半以上均集中在数据集市上。

数据集市不等于数据仓库，多个数据集市简单合并起来不能成为数据仓库。

数据集市的特性：①规模小；②特定的应用；③面向部门；④由业务部门定义、设计和开发；⑤由业务部门管理和维护；⑥快速实现；⑦购买较便宜；⑧投资快速回收；⑨可升级到完整的数据仓库。

数据仓库是企业级的，能为整个企业的运行提供决策支持手段；而数据集市则是部门级

的，一般只能为某个部门内的管理人员服务，因此也称之为部门级数据仓库。

数据集市有两种，即从属的数据集市（Dependent Data Mart）和独立的数据集市（Independent Data Mart）。

1）从属数据集市

从属数据集市的逻辑结构如图 5.2 所示。

所谓从属，是指它的数据直接来自于中央数据仓库。显然，这种结构仍能保持数据的一致性。一般为那些访问数据仓库十分频繁的关键业务部门建立从属的数据集市，这样可以很好地提高查询的反应速度。

2）独立数据集市

独立数据集市的逻辑结构如图 5.3 所示。

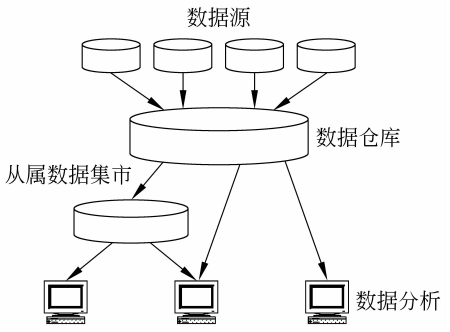

图 5.2　从属数据集市结构

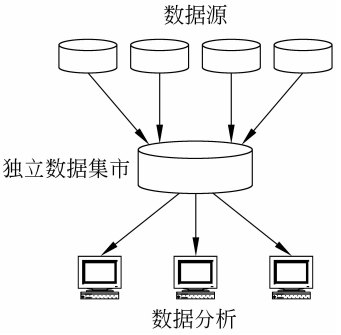

图 5.3　独立数据集市结构

独立数据集市的数据直接来源于各生产系统。许多企业在计划实施数据仓库时，往往出于投资方面的考虑，最后建成独立数据集市，用来解决个别部门比较迫切的决策问题。从这个意义上讲，它和企业数据仓库除了在数据量大小和服务对象上有所区别外，逻辑结构并无多大区别，这是把数据集市称为部门数据仓库的主要原因。

5.1.3　元数据

元数据是数据仓库的重要组成部分。元数据描述了数据仓库的数据和环境，即关于数据的数据（data about data）。元数据可分为 4 类，分别为：关于数据源的元数据、关于数据模型的元数据、关于数据仓库映射的元数据和关于数据仓库使用的元数据。

元数据就相当于数据库系统中的数据字典。由于数据仓库与数据库有很大的不同，因此元数据的作用远不是数据字典所能相比的。元数据在数据仓库中有着举足轻重的作用，它不仅定义了数据仓库有什么，指明了数据仓库中信息的内容和位置，刻画了数据的抽取和转换规则，存储了与数据仓库主题有关的各种商业信息，而且整个数据仓库的运行都是基于元数据的，如数据的修改、跟踪、抽取、装入、综合等。

1. 关于数据源的元数据

它是现有的业务系统的数据源的描述信息。这类元数据是对不同平台上的数据源的物理结构和含义的描述。具体如下。

(1) 数据源中所有的物理数据结构,包括所有的数据项及数据类型。

(2) 所有数据项的业务定义。

2. 关于数据模型的元数据

这类元数据描述了数据仓库中有什么数据以及数据之间的关系,它们是用户管理数据仓库的基础。这类元数据可以支持用户从数据仓库中获取数据。

3. 关于数据仓库映射的元数据

这类元数据是数据源与数据仓库数据之间的映射。

当数据源中的一个数据项与数据仓库建立了映射关系,就应该记下这些数据项发生的任何变换或变动,即用元数据反映数据仓库中的数据项是从哪个特定的数据源来的,经过哪些抽取、转换和加载过程。

从源系统的数据到数据仓库中的目标数据的转移是一项复杂的工作,其工作量占整个数据仓库开发的 70%。

4. 关于数据仓库使用的元数据

这类元数据是数据仓库中信息的使用情况描述。

数据仓库的用户最关心的是两类元数据。

(1) 元数据告诉数据仓库中有什么数据,它们从哪里来,即如何按主题查看数据仓库的内容。

(2) 元数据提供已有的可重复利用的查询语言信息。如果某个查询能够满足用户的需求,或者与用户的愿望相似,用户就可以再次使用那些查询而不必从头开始编程。

关于数据仓库使用的元数据能帮助用户到数据仓库中查询所需要的信息,用于解决企业问题。

5.1.4 数据仓库的存储

数据仓库不同于数据库。数据仓库存储的数据模型是数据的多维视图,它直接影响前端工具和联机分析处理的查询引擎。

在多维数据模型中,一部分数据是数量值,如销售量、投资额、收入等。而这些数量值是依赖于一组“维”的,这些维提供了数量值的上下文关系。例如,销售量与城市、商品名称、销售时间有关,这些相关的维唯一决定了这个销售数量值。因此,多维数据视图就是这些由多个维构成的多维空间中存放着数量值。数据仓库数据的存储示意图如图 5.4 所示,图中的小格内存储的数据可以假设为商品的销售量。

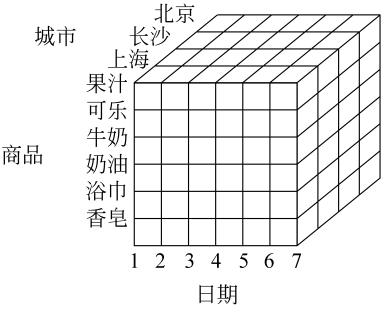

图 5.4 数据仓库数据的存储示意图

多维数据模型的另一个特点是对一个或多个维所做的集合运算。例如,对总销售量按城市进行统计和

排序。这些运算还包括对于同样维所限定的数量值的比较(如销售与预算)。一般来说,时间维是一个有特殊意义的维,它对决策中的趋势分析很重要。

对于逻辑上的多维数据模型,可以使用不同的存储机制和表示模式来实现多维数据模型。目前,使用的多维数据模型主要有星形模型、雪花模型、星网模型等。

1. 星形模型

大多数的数据仓库都采用"星形模型"。星形模型是由"事实表"(大表)以及多个"维表"(小表)所组成。"事实表"中存放大量关于企业的事实数据(数量数据)。通常都很大,而且非规范化程度很高。例如,多个时期的数据可能会出现在同一个表中。"维表"中存放描述性数据,维表是围绕事实表建立的较小的表。

图 5.5 所示的是一个星形数据模型实例。

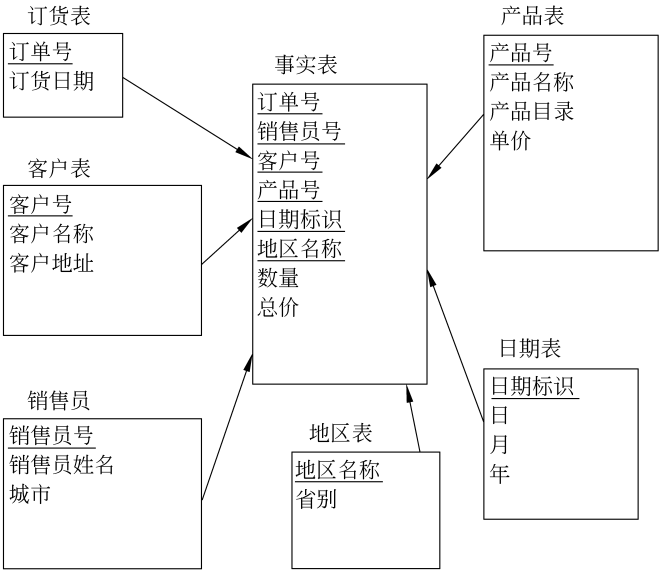

图 5.5　星形数据模型实例

事实表有大量的行(记录),然而维表相对来说有较少的行(记录)。星形模型存取数据速度快,主要在于针对各个维做了大量的预处理,如按照维进行预先的统计、分类、排序等,如按照汽车的型号、颜色、代理商进行预先的销售量统计,作报表时速度会很快。

2. 雪花模型

雪花模型是对星形模型的扩展,雪花模型对星形模型的维表进一步层次化,原来的各维表可能被扩展为小的事实表,形成一些局部的"层次"区域。它的优点是最大限度地减少数据存储量,以及把较小的维表联合在一起来改善查询性能。

在上面星形模型的数据中,对"产品表""日期表""地区表"进行扩展形成雪花模型数据,如图 5.6 所示。使用数据仓库的工具完成一些简单的二维或三维查询,既满足了用户对复杂的数据仓库查询的需求,又能够完成一些简单查询功能而不用访问过多的数据。

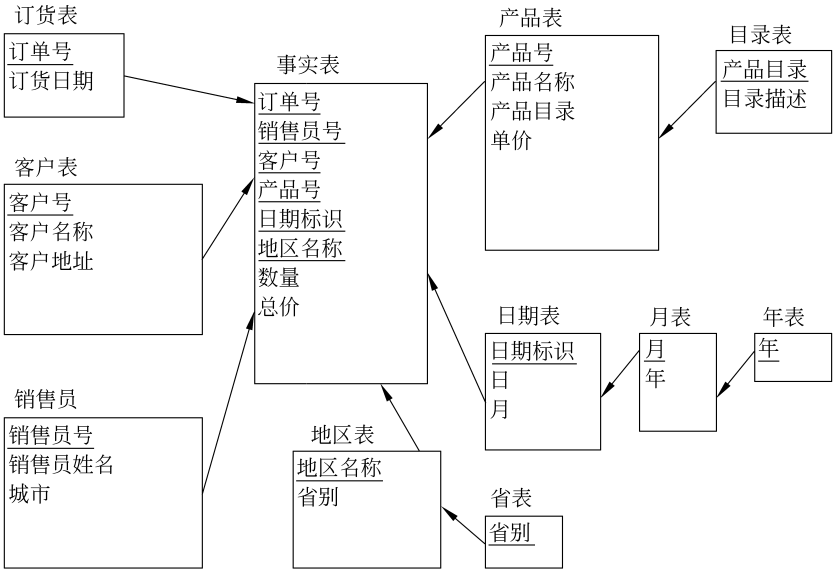

图 5.6　雪花数据模型实例

3. 星网模型

星网模型是将多个星形模型连接起来形成网状结构。多个星形模型通过相同的维，如时间维，连接多个事实表。

5.1.5　数据仓库系统

数据仓库系统由数据仓库、仓库管理和分析工具 3 部分组成，其结构形式如图 5.7 所示。

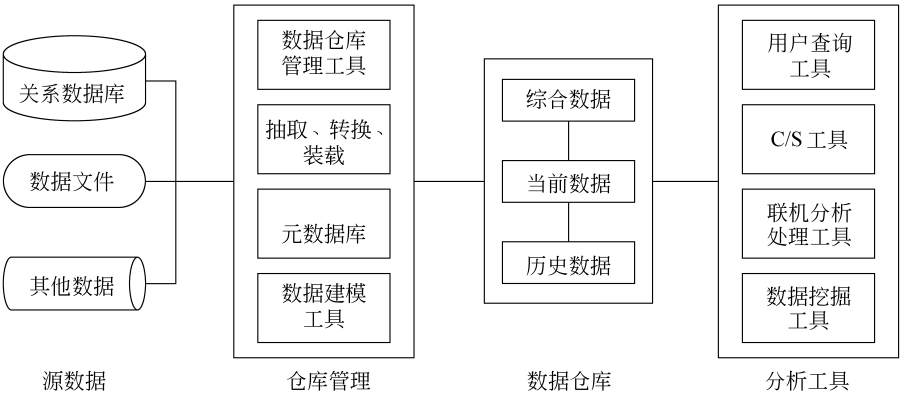

图 5.7　数据仓库系统结构图

数据仓库的数据来源于多个数据源。源数据包括企业内部数据、市场调查报告以及各种文档之类的外部数据。

1. 数据仓库管理系统

在确定数据仓库信息需求之后，首先进行数据建模，确定从源数据到数据仓库的数据抽取、清理和转换过程，划分维数以及确定数据仓库的物理存储结构。元数据是数据仓库的核心，它用于存储数据模型，定义数据结构、转换规划、仓库结构、控制信息等。仓库的管理包括对数据的安全、归档、备份、维护、恢复等工作，这些工作需通过数据仓库管理系统(DWMS)来完成。

数据仓库管理系统由以下几部分组成。

1) 定义部分

定义部分用于定义和建立数据仓库系统。它包括设计和定义数据仓库的数据库、定义数据来源、确定从源数据向数据仓库复制数据时的清理和增强规则。

2) 数据获取部分

该部件把数据从源数据中提取出来，依定义部件的规则，抽取、转换和装载数据进入数据仓库。

3) 管理部分

它用于管理数据仓库的工作，包括对数据仓库中数据的管理、将仓库数据取出给分布的DSS用户、对仓库数据的安全、归档、备份、恢复等处理工作。

4) 目录部分

数据仓库的目录数据是元数据，由以下3方面组成。

(1) 技术目录：由定义部分生成，关于数据源、目标、清理规则、变换规则以及数据源和仓库之间的映像信息。

(2) 业务目录：由仓库管理员生成，包括仓库数据的来源及当前值、预定义的查询和报表细节、合法性要求等。

(3) 信息引导器：使用户容易访问仓库数据。包括查询和引导功能，利用固定查询或建立新的查询，生成暂时的或永久的仓库数据集合的能力等。

该部分是数据仓库使用能力的关键因素。

5) 数据库管理系统部分

数据仓库的存储形式仍为关系型数据库，因此需要利用数据库管理系统。由于数据仓库含大量的数据，要求数据库管理系统产品提供高速性能。

2. 数据仓库工具集

由于数据仓库的数据量大，必须有一套功能很强的分析工具集来实现从数据仓库中提供辅助决策的信息，完成决策支持的各种要求。

分析工具集包括以下两类工具。

1) 查询工具

数据仓库的查询不是指对记录级数据的查询，而是指对分析要求的查询。一般包括以下两类工具。

(1) 可视化工具：以图形化方式展示数据，可以帮助了解数据的结构、关系以及动

态性。

(2) 多维分析工具(OLAP 工具)：通过对信息的多种可能的观察形式进行快速、一致和交互性的存取，这样便于用户对数据进行深入的分析和观察。

多维数据的每一维代表对数据的一个特定的观察视角，如时间、地域、商品等。

2) 挖掘工具

从大量数据中挖掘具有规律性的知识，需要利用数据挖掘(Data Mining)工具。

3. 数据仓库的运行结构

数据仓库应用是一个典型的客户/服务器(C/S)结构形式。数据仓库采用服务器结构，客户端所做的工作有：客户交互、格式化查询、结果显示、报表生成等。服务器端完成各种辅助决策的 SQL 查询、复杂的计算和各类综合功能等。现在，越来越普通的一种形式是 3 层 C/S 结构形式，即在客户与数据仓库服务器之间增加一个多维数据分析服务器，如图 5.8 所示。

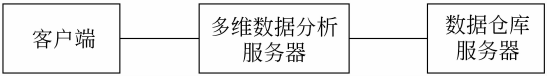

图 5.8　数据仓库应用的 3 层 C/S 结构

多维数据分析服务器将加强和规范化决策支持的服务工作，集中和简化了原客户端和数据仓库服务器的部分工作，降低了系统数据传输量。这种结构形式工作效率更高。

5.2　联机分析处理

联机分析处理(On Line Analytical Processing，OLAP)的概念最早是由关系数据库之父 E. F. Codd 于 1993 年提出的。当时，Codd 认为随着企业数据量的急剧增加，联机事务处理已经不能满足终端用户对数据库查询分析的需要，决策分析需要对关系数据库进行大量的计算才能得到结果，而且查询的结果并不能满足决策者所提出的问题。因此 Codd 提出了多维数据库和多维分析的概念，即多维数据分析的概念。

在数据仓库系统中，联机分析处理是重要的数据分析工具，它的基本思想是企业的决策者应能灵活地操纵企业的数据，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。

5.2.1　基本概念

近十几年来，人们利用信息技术生产和收集数据的能力大幅度提高，大量的数据库被用于商业管理、政府办公、科学研究和工程开发等，这一势头仍将持续发展下去。于是，一个新的挑战被提了出来：在信息爆炸时代，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成为包袱，甚至成为垃圾。联机分析处理是解决这类问题的最有力工具之一。

联机分析处理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人，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。联机分析处理软件以它先进的分析功能和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。

1. 联机分析处理的定义

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。例如，决策者可能想知道“东北地区和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 万元、50 万～100 万元，以及 100 万元以上分组”。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，我们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。

联机分析处理是在联机事务处理的基础上发展起来的，联机事务处理是以数据库为基础的，面对的是操作人员和低层管理人员，在网络上对基本数据的查询和增、删、改等进行处理。而联机分析处理是以数据仓库为基础的数据分析处理。它有两个特点：一是在线联机（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来完成的；二是多维分析（Multi-dimension Analysis），这也是联机分析处理的核心所在。

联机分析处理超越了联机事务处理的查询和报表的功能，它的决策支持能力更强。在多维数据环境中，联机分析处理为终端用户提供了复杂的数据分析功能。通过联机分析处理，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，比较普遍接受的联机分析处理的定义有两种。

1）联机分析处理理事会给出的定义

联机分析处理是一种软件技术，它使分析人员能够迅速、一致、交互地从各个方面（维，即坐标）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（多维）。

企业用户的观点要求数据是多维的。拿销售来说，不仅可从生产这方面看，还与地点、时间等有关，这就是为什么要求联机分析处理模型是多维的原因。这种多维用户视图通过一种更为直观的分析模型进行分析和设计。

联机分析处理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，也被看成超立方体。沿着各个维方向存储数据，它允许用户沿事物的维的要求能方便地分析数据。

2）联机分析处理的简单定义

近年来，随着人们对联机分析处理理解的不断深入，有些学者提出了更为简要的定义，