第三讲 数据库与数据库管理

【教学目的和要求】

1. 理解数据库在人工管理、文件管理、数据库系统的三个阶段的发展过程和特点
2. 数据库系统的体系结构
3. 掌握从现实世界到概念模型和数据模型抽象的含义，不同的数据库模型，实体关系图画法，
4. 理解数据库管理系统（DBMS）的功能及其工作过程
5. 了解多媒体数据库的组成
6. 了解数据仓库和数据挖掘的概念

【主要内容】

【电子教案】

参见：第三讲数据库与数据库系统

【重点与难点】

1. 数据库在人工管理、文件管理、数据库系统的三个阶段的发展过程
2. 数据库系统的体系结构；
3. 数据模型。

【教材和参考读物】

《管理信息系统——理论与实践》第三章

《管理信息系统》甘仞初 第三章

【教学时数】 2

第 3 讲 数据库与信息管理

数据库技术是计算机科学的一个重要分支。20 世纪 50 年代以来，计算机应用由科学研究逐步扩展到企业、政府部门和社会的各个领域，数据处理很快上升为计算机应用的一个最重要的方面。自 1968 年第一个商品化数据管理系统问世以来，数据库技术得到迅速发展。近年来，随着网络技术和多媒体技术的发展，基于互联网的融合多媒体技术的数据库技术显示出更为广阔的技术前景，成为信息管理、办公自动化的主要技术支持手段。

数据库技术研究如何科学地组织数据和存储数据，如何高效地检索数据和处理数据，以及如何既减少数据冗余，又能保障数据安全，实现数据共享。在计算机应用的领域中，管理信息系统方面的应用占 90%以上，而数据库技术又是管理信息系统的基础。因此，可以说，数据库是当今计算机应用中覆盖范围最为广泛的应用。

3.1 数据库与数据库管理系统

3.1.1 数据库技术的发展

数据处理的首要任务是数据管理。数据管理是指如何分类、组织、存储、检索及维护数据库。数据管理技术经历了人工管理、文件管理、数据库系统三个阶段。表 3-1 给出了三个阶段的特征比较。

表 3-1 数据管理的三个阶段

	人工管理	文件管理	数据库管理
数据管理者	用户	文件系统	数据库管理系统
面向的对象	某个应用程序	某个应用	多个应用
共享程度	无共享，冗余度大	共享性差，冗余度大	共享性好，冗余度小
独立性	不独立，与程序一体化	独立性差，与程序相关性强	具有高度的物理独立性和逻辑独立性
结构化	无结构	文件形式多样化，单个文件有记录结构，文件之间是独立的	整体结构化程度高，以数据模型描述
控制	应用程序自己控制	应用程序自己控制	数据库管理系统提供安全性、完整性、并发控制和恢复能力

1．人工管理阶段

从 1946 年计算机诞生至 20 世纪 50 年代中期，计算机主要用于科学计算。计算机除硬件设备外没有任何软件可用，使用的外存只有磁带、卡片和纸带，没有磁盘等直接存取设备。软件中只有汇编语言，没有操作系统，对数据的处理，完全由人工进行管理。

人工管理阶段的数据模型如图 3-1 所示。图中显示程序和数据是一体化的，虽然以虚线将程序和数据分成两部分，事实上，它们之间是混为一体的。

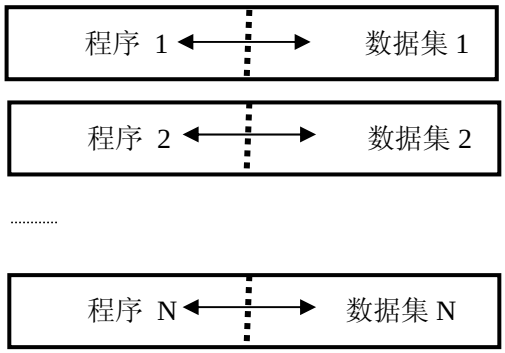


图 3-1 数据人工管理模型

在人工管理阶段，数据管理呈现如下特点：

- 数据不保存。一组数据对应于一个应用程序，应用程序与其处理的数据结合成一个整体。在进行计算时，系统将应用程序和数据一起装入，程序运行结束后，释放内存空间，程序和数据同时被撤销。
- 没有软件对数据进行管理。应用程序设计者不仅要考虑数据之间的逻辑关系，还要考虑存储结构、存取方法以及输入方式等。如果存储结构发生变化，程序中读写数据的程序也要发生改变，数据没有独立性。
- 没有文件概念。数据的组织方法由程序设计人员自行设计和安排。
- 数据面向应用。数据附属于程序，即使两个应用程序使用相同的数据，也必须各自定义数据的存储和存取方式，不能共享相同的数据定义，因此，程序与程序之间可能有大量的重复数据。

2．文件管理阶段

20 世纪 50 年代后期到 60 年代中期，计算机不仅用于科学计算，也大量用于经营管理活动。硬件设备有了磁盘、磁鼓等直接存储设备；软件发展了操作系统和各种高级语言。

文件系统的模型如图 3-2 所示。通过文件系统，程序和数据之间有了比较清晰的边界。不同的程序可以使用相同的文件，反过来，一个程序也可以访问不同的文件。

文件系统阶段数据管理有如下特点：

- 数据可长期保存在磁盘上。用户可通过程序对文件进行查询、修改、插入或删除操作。
- 文件系统提供程序和数据之间的读写方法。文件管理系统是应用程序与数据文件之间的一个接口。应用程序通过文件管理系统建立和存储文件；反之，应用程序要存取文件中的数据，必须通过文件管理系统实现。用户不必关心数据的物理位置，程序和数据之间有了一定的独立性。
- 文件形式多样化。因为有了直接存取设备，所以可以建立索引文件、链接文件和直接存取文件等。对文件的记录可顺序访问和随机访问。文件之间是相互独立的，文件与文件之间的联系需要用程序实现。
- 数据的存取基本上以记录为单位。

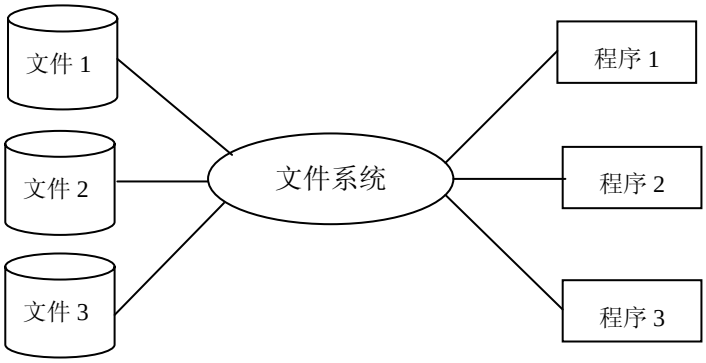


图 3-2 文件系统模型

文件系统的缺陷是：

- 数据冗余大，因为文件是为特定的用途设计的，因此会造成数据在多个文件中重复存储。
- 数据的不一致。这是由数据冗余和文件之间的独立性造成，在更新数据时，很难保证同一数据在不同文件中的统一。
- 程序与数据之间的独立性差。修改文件的存储结构后，相关的程序也要修改。

3．数据库管理阶段

20 世纪 60 年代后期开始，存储技术有了很大的发展，产生了大容量磁盘。计算机用于管理的规模更加庞大，数据量急剧增长，为了提高效率，人们着手开发和研制更加有效的数据管理模式，并由此提出了数据库的概念。

1968 年，IBM 公司研制成功数据库管理系统（Information Management System，IMS）标志着数据管理技术进入了数据库阶段。IMS 为层次型数据库。1969 年，美国数据系统语言协会（Conference On Data System Language）公布了数据库工作组报告，对研制开发网状数据库起了巨大推动作用。1970 年，IBM 公司的研究员 E F. Codd 连续发表论文，奠定了关系数据库的基础。

数据库系统的数据存取模型如图 3-3 所示。

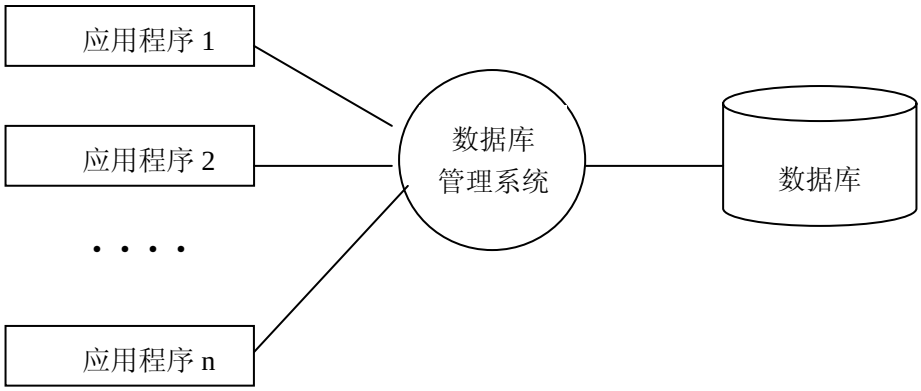


图 3-3 数据共享示意图

与文件管理相比，数据库技术有了很大的改进，主要表现为：

- 数据库中的数据是结构化的。在文件系统中，数据是无结构的，即不同文件中的记录之间

没有联系，它只在数据项之间有联系。数据库系统不仅考虑数据项之间的联系，还要考虑记录之间的联系，这种联系是通过存储路径来实现的。

- 数据库中的数据是面向系统的，对于任何一个系统来说，数据库中的数据结构是透明的。任何应用程序都可以通过标准化接口访问数据库，如图 3-3 所示。
- 数据库系统比文件系统有较高的数据独立性。
- 数据库系统为用户提供了方便统一的接口。用户可以用数据库系统提供的查询语言和交互式命令操纵数据库。用户也可以用高级语言编写程序来访问数据库，扩展了数据库的应用范围。

不仅如此，数据库技术的发展使数据管理上了一个新台阶，在数据完整性、安全性、并发访问和数据恢复方面，数据库管理系统都提供了非常完善的功能选择。

- 数据完整性

保证数据库存储数据的正确性。例如预定同一班飞机的旅客不能超过飞机的定员数；定购货物中，订货日期不能大于发货日期。使用数据库系统提供的存取方法，设计一些完整性规则，对数据值之间的联系进行校验，可以保证数据库中数据的正确性。

- 数据安全性

并非每个应用都可以存取数据库中的全部数据。例如在一个人事档案数据库中，只有被授权的访问者才可以读取数据，并进行修改；其他访问者的权限一般限于浏览特定的数据项，而不是全部数据。

- 并发控制

当多个用户同时存取、修改数据库中的数据时，可能会发生相互干扰，使数据库中的数据完整性受到破坏，而导致数据的不一致。数据库并发控制防止了这种现象的发生，提高了数据库的访问效率。

- 数据库的恢复

任何系统都不可能永远正确无误地工作，数据库系统也是如此。运行过程中，会出现硬件或软件的故障。数据库系统具有恢复能力，能把数据库恢复到最近某个时刻的正确状态。

3.1.2 数据库系统的体系结构

可以从不同的角度分析数据库系统的体系结构，从 DBMS 角度看，数据库系统采用三级模式结构，也就是内模式、外模式和概念模式；从数据库的物理分布来考察，又分为集中式数据库、C/S 结构、B/S 结构等，这就是数据库系统的体系结构。

目前市场上流行的数据库系统软件产品多种多样，支持不同的数据模型，使用不同的数据库语言和应用系统开发工具，建立在不同的操作系统之上，但绝大多数数据库都具有三级模式的特征。数据库的三级模式分为：外模式、内模式和概念模式，如图 3-4 所示。

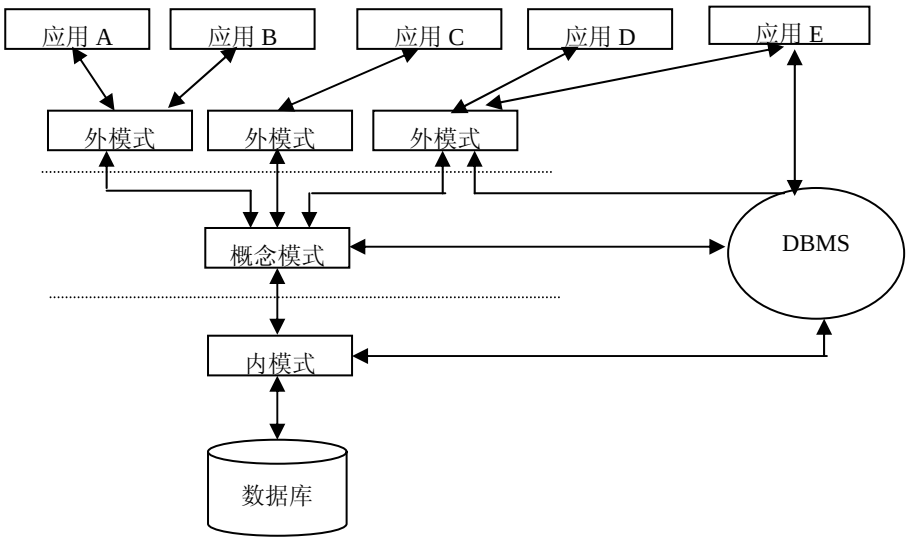


图 3-4　数据库三级模式

外模式定义了允许用户操作的数据库数据，也称为用户模式或子模式。对最终用户来讲，所看到的视图就是外模式。由于不同用户需求相差很大，看待数据的方式与所使用的数据内容各不相同，对数据的保密性要求也各有差异，因此，不同用户的外模式也不相同。

概念模式，简称为模式，是数据库全部数据的逻辑结构和特征描述，它以数据模型为基础，采用数据库系统提供的模式描述语言进行定义，可以被看作是现实世界中一个组织或部门中实体及其联系的抽象模型在数据库系统中的实现。概念模式不同于外模式，与具体的应用程序无关；也不同于内模式，与数据库的硬件环境与存储格式无关。

概念模式不仅要定义数据的逻辑结构，而且要定义与数据有关的安全性和完整性；不仅要定义数据记录的内部结构，还要定义这些数据之间的联系。

内模式也称为存储模式，用来描述数据的物理结构和存储方式。

数据库三级模式的意义在于提供数据的层次结构，保持数据的独立性。内模式到概念模式之间的分割提供了数据的物理独立性，即当数据的物理结构发生变化时，如存储设备的改变、数据存储位置或存储组织方式的改变等，不影响数据的逻辑结构。例如，为了提高数据的存取效率，数据库设计人员重新组织数据的物理组织，这种改变由于内模式与概念模式的存在，而使得数据的概念模式不会受到影响，也不需要修改应用程序。

概念模式到外模式的映像提供了数据的逻辑独立性，即当数据的整体逻辑结构发生变化时，如为原有记录增加新的数据项、在概念模式中增加新的数据类型、增加新的数据库记录等，都不影响外模式。例如，在采购系统中，因为产品结构的变化，采购的零部件需要增、删、修改、更新等，根据新的数据需求修改数据库之后，并不引起应用程序的变化。

数据库的三级模式，提供了高度的数据独立性。其中，数据库的全局逻辑描述是独立于其他所有结构描述的，在定义数据库结构时，应该首先定义概念模式。内模式则是将概念模式中所定义的数据进行适当的组织并加以存储，以实现较好的时空效率。

总之，数据库的三级模式是数据库管理的结构框架，依照这些数据框架组织的数据才是数据库内容。在数据库设计时，主要是定义数据库的三级模式，而在用户使用数据库时，关心的是数据库的内容。数据库的模式通常是稳定的，而数据库的数据通常是经常变化的，特别是来自企业业务流程的数据，数据始终处于动态变化之中。

3.1.4 数据库管理系统（DBMS）的功能及其工作过程

1．数据库管理系统的主要功能

（1） 数据库的定义功能

DBMS 提供数据描述语言（DDL），定义数据库的外模式、概念模式、内模式、数据的完整性约束和用户的权限等。例如 Oracle 的数据库管理系统提供 DDL，定义 Oracle 数据库的表、视图、索引等各种对象。DBMS 把用 DDL 写的各种源模式翻译成内部模式，放在数据字典中，作为管理和存取数据的依据。例如 DBMS 可把应用的查询请求从外模式，通过模式转化到物理记录，查询出结果返回给应用。

（2） 数据操纵功能

DBMS 提供的数据操纵语言（Data Manipulation Language，DML）可实现对数据的插入、删除和修改等操作。DML 语言有两种用法：一种方法是把 DML 语句嵌入到高级语言中，另一种方法是交互式地使用 DML 语句。对于第一种方法，DBMS 必须提供预编译程序，预处理嵌入 DML 语句的源程序，识别 DML 语句，转换为相应高级语言能调用的语句，以便原来的编译程序能接受和处理它们。

（3） 数据库的控制功能

数据库的控制功能包括并发控制、数据的安全性控制、数据的完备性控制和权限控制，保证数据库系统的正确有效运行。

（4） 数据库的维护功能

已经建立好的数据库，在运行过程中需要进行维护。维护功能包括数据库出现故障后的恢复、数据库的重组、性能的监视等。这些功能大部分由实用程序来完成。

（5） 数据字典

数据字典（Data Dictionary，DD）中存放着数据库体系结构的描述。对于应用的操作，DBMS 都要通过查阅数据字典进行。例如 Oracle 数据库系统，其数据字典中存放着用户建立的表和索引、系统建立的表和索引以及用于恢复数据库的信息等。当增加表、删除表或修改表的内容时，DBMS 自动更新数据字典；当应用检索数据时，Oracle 的 DBMS 动态地将数据字典与用户程序或终端操作连起来，保持系统正确地运行。Access 数据库管理系统动态地提供了对象浏览器，将数据字典以对象的形式同其他数据库对象一起进行管理。

2．数据库管理系统的工作过程

一个数据库系统的建立是按模式和存储模式描述的框架，将原始数据存储到设备介质上形成的。用户可以通过应用程序或查询语言实现对数据的操作。

下面我们以应用程序读取一个记录为例讨论一下 DBMS 的工作过程，以了解 DBMS 与应用程序、操作系统的接口以及三级模式的使用，如图 3-10 所示。

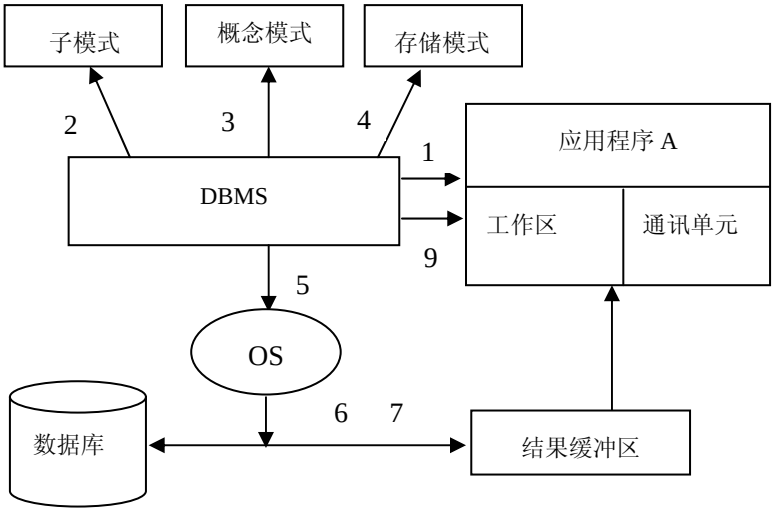


图 3-10 DBMS 工作过程示意图

- 应用程序 A 通过 DML 命令向 DBMS 发出读请求，并提供读取记录参数，如记录号、关键字等。
- DBMS 根据应用程序 A 对应的子模式中的信息，检查用户权限，决定是否接受读请求。
- 如果是合法用户，则调用模式，根据模式与子模式间数据的对应关系，确定需要读取的逻辑数据记录。
- DBMS 根据存储模式，确定需要读取得物理记录。
- DBMS 向操作系统发读取记录的命令。
- 操作系统执行该命令，控制存储设备读出记录数据。
- 在操作系统控制下，将读出的记录送入系统缓冲区。
- DBMS 比较模式与子模式，从系统缓冲区中得到所需的逻辑记录，并经过必要的数据变换后，将数据送入用户工作区。
- DBMS 向应用程序发送读命令执行情况的状态信息。
- 应用程序对读取的数据进行相应处理。

3．数据库系统的不同视图

数据库系统的管理、开发和使用人员主要有数据库管理员、系统分析员、应用程序员和用户。这些人员的职责和作用是不同的，因而涉及到不同的数据抽象级别，分别对应于不同的数据视图。如图 3-11 所示。

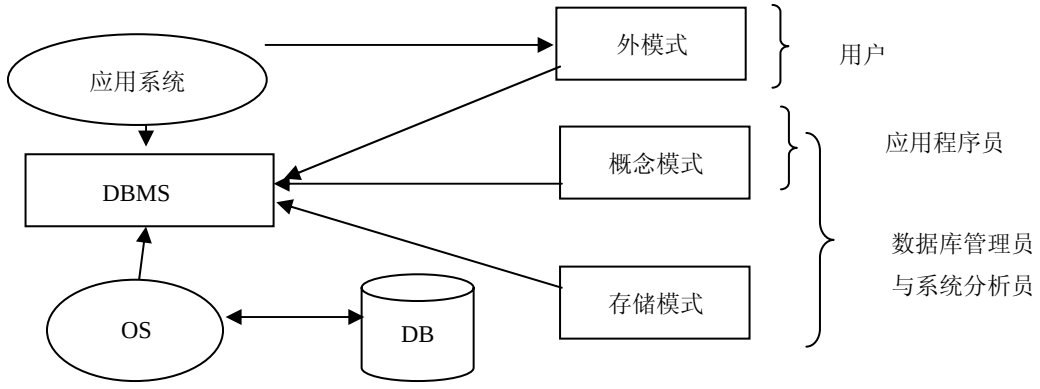


图 3-11　数据库系统的不同视图

（1）用户

用户分为应用程序和最终用户两类（End User），他们通过数据库系统提供的接口和开发工具软件使用数据库。目前常用的接口方式有菜单驱动、表格操作、利用数据库与高级语言的接口编程、生成报表等。这些接口给用户带来很大方便。

（2）应用程序员

应用程序员负责涉及应用系统的的程序接口，编写应用程序通过数据库管理员为他建立的外模式来操纵数据库中的数据。

（3）系统分析员

系统程序员负责应用系统的需求分析和规范说明。系统分析员要与用户和数据库管理员配合好，确定系统的软硬件配置，共同做好数据库各级模式的概要设计。

（4）数据库管理员

数据库管理员（DataBase Administrater，DBA）可以是一个人，也可以是由几个人组成的小组。他们全面负责管理、维护和控制数据库系统，一般来说由业务水平较高和资历较深的人员担任。

他们的主要工作包括：

决定数据库的信息内容。数据库中存放什么信息是由 DBA 决定的。他们确定应用程序的实体，完成数据库模式的设计，并同应用程序员一起完成用户子模式的设计工作。

决定数据库的存储结构和存取策略。确定数据的物理组织、存放方式及数据存取方法。

定义存取权限和有效性检验。用户对数据库的存取权限、数据的保密级别和数据的约束条件都是由 DBA 确定的。

建立数据库。DBA 负责原始数据的装入，建立用户数据库。

监督数据库的运行。DBA 负责监视数据库的正常运行，当出现软硬件故障时，能及时排除，使数据库恢复到正常状态，并负责数据库的定期转储和日志文件的维护等工作。

重组和改进数据库。DBA 通过各种日志和统计数字分析系统性能。当系统性能下降时，对数据库进行重新组织，同时根据用户的使用情况，不断改进数据库的设计，以提高系统性能，满足用户需要。

3.2　数据仓库和数据挖掘

数据仓库是信息技术领域和企业界最新最热门的流行词汇和概念之一。提高顾客满意度，不断增加市场份额和利润，增强企业的市场竞争力等，所有战略性并与企业历史信息相关的重大决策都需要数据仓库技术的支持。数据仓库是信息的逻辑集合，这些信息来自许多不同的业务数据库，并用于支持企业的分析活动和决策任务，或者说，数据仓库代表了一种对企业中的信息进行组织和管理的与以往不同的思维方式。

3.2.1 数据仓库的概念

目前，数据仓库一词尚没有一个统一的定义，著名的数据仓库专家 W.H.Inmon 在其著作《Building the Data Warehouse》一书中给予如下描述：数据仓库（Data Warehouse）是一个面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集

合，用于支持管理决策。对于数据仓库的概念我们可以从两个层次予以理解，首先，数据仓库用于支持决策，面向分析型数据处理，它不同于企业现有的操作型数据库；其次，数据仓库是对多个异构的数据源有效集成，集成后按照主题进行了重组，并包含历史数据，而且存放在数据仓库中的数据一般不再修改。

根据数据仓库概念的含义，数据仓库拥有以下四个特点：

1．面向主题

操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织。主题是一个抽象的概念，是指用户使用数据仓库进行决策时所关心的重点方面，一个主题通常与多个操作型信息系统相关。

2．集成的

面向事务处理的操作型数据库通常与某些特定的应用相关，数据库之间相互独立，并且往往是异构的。而数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

3．相对稳定的

操作型数据库中的数据通常实时更新，数据根据需要及时发生变化。数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

4．反映历史变化

操作型数据库主要关心当前某一个时间段内的数据，而数据仓库中的数据通常包含历史信息，系统记录了企业从过去某一时点(如开始应用数据仓库的时点)到目前的各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测。

企业数据仓库的建设，是以现有企业业务系统和大量业务数据的积累为基础。数据仓库不是静态的概念，只有把信息及时交给需要这些信息的使用者，供他们做出改善其业务经营的决策，信息才能发挥作用，信息才有意义。而把信息加以整理归纳和重组，并及时提供给相应的管理决策人员，是数据仓库的根本任务。因此，从产业界的角度看，数据仓库建设是一个工程，是一个过程。

3.2.2 为什么需要数据仓库

商业活动的复杂性以及顾客对企业响应速度越来越苛刻的需求，改变了企业的经营方式。企业经理人员不仅要了解市场发生的事情，还要知道为什么会发生这些事情。而为了回答“为什么会发生”，数据仓库技术可以起到关键作用。数据仓库在综合各种业务数据的基础上，以多维数据库和数据挖掘为工具，提供智能查询和大量的总结报告。

1．管理中的问题

新世纪商业环境的一大特征是外部力量加剧了市场竞争，企业必须寻求市场差异性，或者支持更快的响应速度。企业的历史数据是一种极其重要的信息，它与顾客、顾客/产品关系、顾客购买模式等有关。数据仓库具有将信息转换成知识的潜在能力，顾客的深层次信息可以潜在地传送给经理。

图 3-12 所示的模型说明了数据仓库在顾客分类和市场竞争方面的应用。

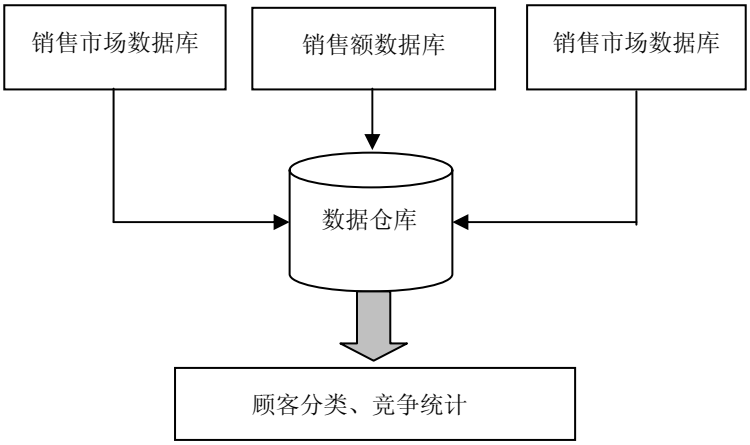


图 3-12 利用数据仓库获得竞争性信息

在图 3-12 中，数据仓库将企业各个业务数据库中的信息结合起来（通过汇总和合计）。当人们从各类业务数据库中提取信息来创建数据库时，收集的只是那些进行决策所需的信息。比如 MasterCard 公司和它的数据仓库，以及 MasterCard 的联机系统，就提供了这方面的支持。该公司的数据仓库除了其是世界上最大的数据仓库这一事实之外，还可以为与它合作的银行、商店、饭店等合作伙伴挖掘有价值的信息。如果你正在为一家饭店工作，并想以赠机票的方式作为促销活动的一部分，那么 MasterCard 的数据仓库可以为顾客建立这样的联机分析查询："那些经常最少一月两次来往我们饭店的顾客，他们喜欢去的目的地是哪儿？"

2．现有系统的现状

构造数据仓库的另外一方面的动力源于现有系统不合适，以及缺少商业信息。许多产品系统无法满足商业用户的需要。通常，无论在形式上还是在实际上都不可访问数据，而且数据也是不一致的。由于数据不一致，例如不同报表中的销售数据无法匹配，使得商业用户不能精确描述其收入。缺少通用尺度意味着决策者无法认清市场形势并正确地评估企业的经营行为。

市场和销售人员需要尽快地访问数据，较快较多地获得报表，快速地分析，以及做出较为敏锐的反应，以便管理商业活动并增加收入。甚至在创建和生成报表时，实际的信息是过时的。

不同的产品系统对不同数据库中同一顾客保存不同的信息。这样无法按统一且完整的方式来看待每一位顾客，可能会导致在交叉购买、目标市场、产品包装等方面丧失机会。如果顾客需要一次购买所有东西，而不是在同一公司的不同人处购买，他们无法获得服务。这样，商业性能不是有所增长，而是有所下降。

3.2.3 数据仓库的价值

数据仓库的潜在价值很大，而且可能进一步增大，这些价值包括三方面的内容：

1．成本/效率决策支持

数据仓库可从产品系统中下载报表和即时查询。此外，商业用户不再需要信息技术专家的支持。当提供集成的、简洁的、一致的数据时，有利于增加报表和查询的质量和可靠性。

决策支持服务包括信息处理、分析处理和数据挖掘，用它们来获得基于实际数据的可采取行动的建议。完整理解全局顾客关系有助于加强对顾客的服务，并最终使顾客变得满意。逻辑集成有助于财务经理更好地管理资产，从而可降低资产的损耗并提高资产的回报率。加深对顾客购买模式的

理解有助于在恰当的时间和地点进行恰当存货，这将进一步提高资产回报率。

2．重组应用系统

将产品系统和决策支持数据仓库分开，这可使信息技术在产品生命期中清除历史数据，并改进企业系统结构。此种清除可以增加产品系统的生命力，或推迟或消除升级的需求。它要求用干净并且一致的数据来加载数据仓库，利用反馈可提高产品系统中数据的质量。对于有些企业，数据仓库可能正是它们所期盼的客户/服务器体系。

3．重构业务流程

数据仓库也可以评价商业和组织的竞争力。外部数据集成提供了评价和分析竞争力的恰当标准。由于数据仓库事实上主要用于了解为什么发生商业事件，而不是发生了什么事件，设计并使用数据仓库有助于管理者了解企业的商业特性。了解商业事件的原因和形式有助于弄清楚是什么力量促使商业人员所追求的战略性益处。

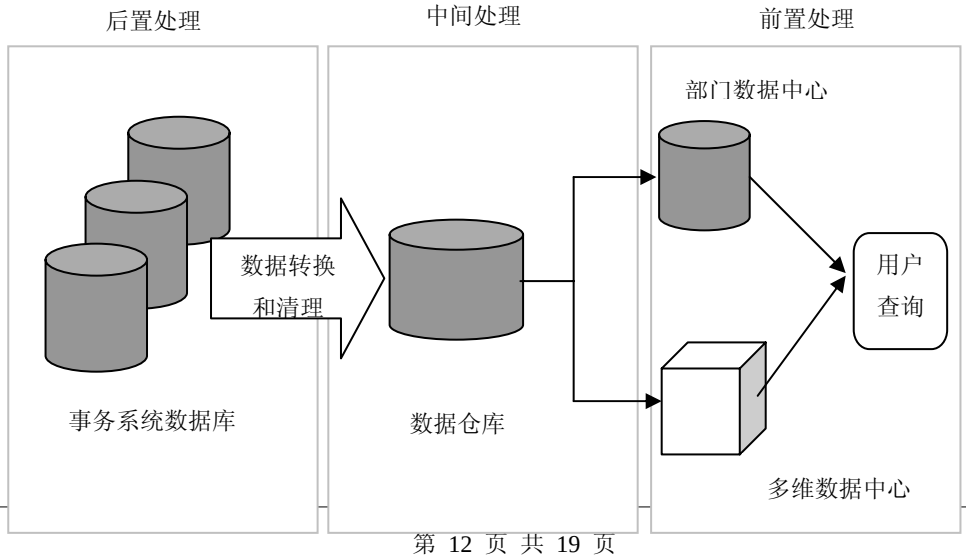
重构业务流程带来的潜在利益比使用成本/效益率决策大得多。再组织应用系统不仅增强了资产的使用，并且对另两方面价值的实现也是必须的。实际上，这三方面是紧密集成的。使用数据仓库的经验表明，当数据仓库从决策支持变成支持业务流程重构时，商业盈利增多，因为人们的效率、资产以及整个企业的能力都有所增加。

3.2.4 数据仓库框架结构

数据仓库提供了一种使信息可用于决策制定的方法。一个有效的数据仓库战略必须能处理现代企业的复杂事件。每个事务处理系统都会产生数据，并且存放在不同的数据库中。用户需要随时随地的访问数据，以满足他们对数据的需求。因此，一个数据仓库必须适应商业模式，而不是支配和改变它。

数据仓库使得经理、管理人员、分析专家和用户能够从它们业务活动中许多方面查询和分析公司数据。数据仓库允许用户对下列数据进行复杂的分析：为适应特别指标而摘录的、聚集的和汇总的数据，被操作用来获取新数据的数据、为取消不想要和不必要的数据而重新格式化或过滤的数据，和其他数据源集成在一起的数据。

许多模块构成了数据仓库系统。这个系统从现有的事务处理系统和 MIS 为基础，一部分为支持数据仓库而设的后台处理，以访问和运用数据仓库内数据的用户而结束。在中间是个分散过程，它使数据以一种局部而不是集中的方式支持用户。至于其他系统，则是覆盖这些处理过程技术的基础，如安全系统，它不仅控制着在终端数据仓库的输入过程，还控制着用户在数据仓库的前台访问能力。数据仓库的框架结构如图 3-13 所示。



安全处理

图 3-13 数据仓库框架结构

1．后台处理

数据仓库系统的后台处理利用了事务处理系统的数据库，这个处理包括以下两个部分：

- 数据收集 为数据仓库收集数据的过程是从当前事务处理系统开始的。该数据仓库的后台处理需要被分成可管理的几个处理模块。事务处理系统生成必须处理和输入到数据仓库的数据。在数据仓库系统的结构内必须有一种方法来截取和收集那些在事务处理系统内已经改变的数据，主要用于数据仓库的数据处理。
- 数据采集 在收集到事务处理系统数据的变化后，数据仓库的后台处理必须采集所有同以前收集的事务相关的数据。数据采集过程通常仅仅获取驱动数据采集过程的关键信息。

后台处理把数据制备成事务数据库，并用它来更新和供给数据仓库系统。这个过程在整个数据仓库系统中是最复杂的，因为用户正处理多种遗留数据源。这些数据源中的一些较为容易处理，而大部分则不是这样。

2．中间处理

数据仓库系统的中间处理包括以下几个部分：

- 数据清理 在收集到所有事务处理系统数据库的数据后，数据必须在放入数据仓库之前进行清理，以获得一个适当的统一格式和定义。
- 数据的放置和分发 当完成数据清理后，数据就必须放置到数据仓库中——有的在中间，有的在较远的位置或处于两个位置之间。
- 标准报表的编译和索引 在数据已放入数据仓库数据存储器之后，对包含于数据仓库系统内的报表必须进行编译和索引。在这个过程结束后，报表很像数据仓库内的原始数据，将让用户在线使用，不必用纸张的形式发送。

3．前台处理

前台处理涉及允许用户对数据仓库所包含的信息进行正确的访问，及提供用户工具集所需的目录和中间数据信息。大多数数据仓库项目的目标应当是驱使这一过程进入强大的用户领域，并脱离信息系统空间。然而，需要构造几个关键的应用程序以用于经验不足的数据仓库用户。这个处理同传统应用程序开发过程很相似。该过程的任务包括用新的信息内容来更新访问数据仓库的应用程序，通过适当的用户工具组内的视图或分类定义来提高访问能力。例如，前台应用程序和整个处理可以给用户提供元数据，这些数据通过最新的处理来告知用户，数据仓库提供的金融数据是正确的。

3.2.5 数据挖掘

1．数据挖掘的定义

数据挖掘也可以称为知识发现（Knowledge Discovery in Database），是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的，人们事先难以预计的，潜在的有价值信息和知识的过程。

所谓知识，并没有一个完整和精确的定义，在此包括数据、信息、概念、规则、模式、规律和约束等。通常人们将数据看成知识的源泉，就像沙里淘金一样。原始数据必须是大量的、来自于现

实的数据。发现的知识可用于信息管理、查询优化、决策支持和过程控制，还可以用于数据自身的维护。因此，数据挖掘是一门交叉学科，它把数据应用从低层次的简单查询，提升到挖掘知识和决策支持。

在数据挖掘中发现的知识，并不是崭新的自然科学定理或者数学公式，而是数据之间存在的某一种关联。这种关联对不同的人呈现出完全不同的价值，比如，购买果酱的人60%同时购买了面包，对超市经营者是非常难得的商业信息，而对消费者而言，几乎没有任何意义。数据挖掘发现的知识都是基于现实当中产生的数据，在特定前提和约束条件下，面向特定领域的知识。用这样的知识可以指导一定范围内的业务活动。知识的表现也要求易于被用户理解，最好能用自然语言表达所发现的结果。

2．数据挖掘的目标和基本特征

许多单位和组织在耗费了巨额资金建立了规模庞大，覆盖整个企业所有经济活动的数据仓库之后，仍然被一个基本的问题所困扰：如何把握顾客的消费倾向，跟踪客户需求并提高产品的市场份额和市场竞争力。

为达到这样的商业目标，数据挖掘可以帮助用户处理大量的数据，以期在数据仓库中得到“意外”的发现，这些发现是潜在带来更高利润的顾客，而不是任意的新顾客；这些发现是战略性的和富有竞争性的，对企业的未来有方向性的指引。

企业的数据分析划分为三个不同的层次和范畴。首先要了解企业经营活动中发生了什么？其次，要了解为什么会发生，然后依据原因确定企业可以做什么，不可以做什么。图 3-14 描述了数据分析的三个层次。传统的查询、报表和多维分析技术主要集中处理发生了什么，但却很少考虑原因。数据发掘的贡献在于支持最高层次的数据分析，并预测可能采取的行动。

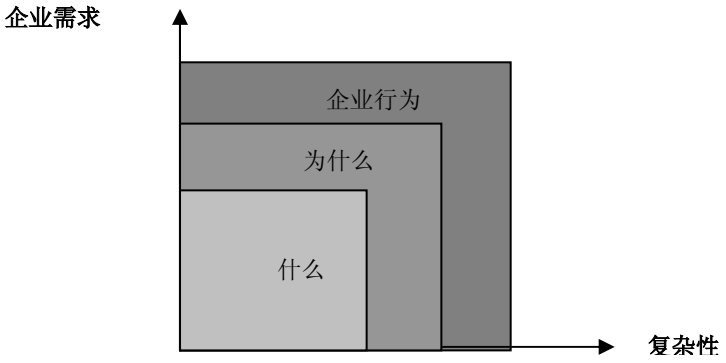


图 3-14 数据分析的层次

数据挖掘工具基于企业已经创建的数据仓库，直接或间接地访问数据源。数据仓库或数据集市的数据经过求精、集成和标准化，服务于数据挖掘工具。从手段上看，数据挖掘与信息和分析处理在很多方面是不同的。表 3-5 总结了数据挖掘的特点。

表 3-5 数据挖掘的特点

	信息分析与处理	数据挖掘
关注点	概括数据	事务数据或细节数据
维数	有限的	许多
属性数目	几十个	每一维都有几百个
数据集的尺寸	每一维都是小型的或中型的	每一维都上百万
分析关注点	商业中会发生什么？	为什么会发生？预测行为

分析技术	片分	自动发现
分析处理	由商业分析员启动和控制	由数据和系统启动所需，商业分析员指导很少
置信因子	由商业分析员得出	由数据生成
技术状况	成熟	统计分析已经成熟，出现了知识发现

3．数据挖掘过程

数据挖掘是一个复杂的过程。数据挖掘充分利用人工智能、机器学习、统计学等多学科的知识，并把他们同其他辅助技术结合到一起，从大量的数据中找出潜在的、有用的知识。数据挖掘的过程如图 3-15 所示。

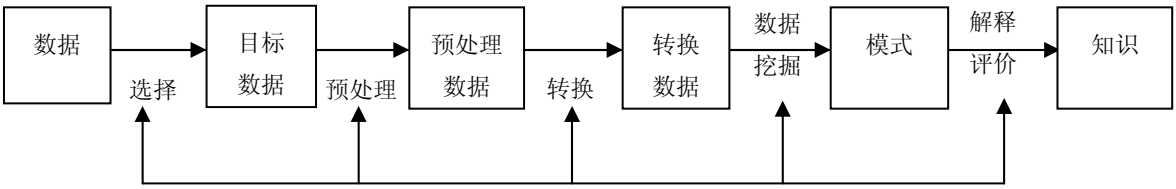


图 3-15　数据挖掘过程

（1） 确定业务对象

首先要清晰地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结果是不可预测的，但要探索的问题应是有预见的，为了数据挖掘而数据挖掘带有盲目性，是不会成功的。

（2） 数据的选择

搜索所由于业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据。进行数据挖掘时，首先要从大量数据中取出一个问题相关的样板数据子集，而不是使用全部数据。通过对数据的取样，选择与知识发现任务相关的数据集，从而减少数据处理量，同时又不降低知识发现的精确度。

（3） 数据的预处理

数据预处理主要完成数据集成和数据清理工作。用于知识发现的数据往往来自于多个实际的系统，存在着异构数据的转换问题、多个数据源的数据之间的冲突，如在命名、结构、取值单位、含义等方面的不同。数据集成是对数据进行统一化和规范化的复杂过程，把原始数据在最底层上加以转换、提炼和集聚，形成最原始的用于知识发现的统一的数据集合。数据清理主要是去除源数据集中的噪声数据和无关数据，处理遗漏数据和清洗冗余数据，考虑数据的时间顺序和数据变化，防止杂乱无章的数据进入数据挖掘阶段。

（4） 数据的转换

将数据转换成一个针对挖掘算法建立的分析模型，找到数据的特征表示。经常用多维数据立方（Data Cube）来组织数据，采用数据仓库中的切换、旋转和投影技术，把初始数据按照不同的层次、粒度和维度进行抽象和聚集，从而生成在不同抽象级别上的知识基。有些数据属性对发现任务是没有影响的，甚至在加入后会大大影响挖掘效率和结果的精确性。数据简化就是在对发现任务和数据本身内容理解的基础上，寻找对发现目标有用的数据的特征，以缩减数据规模，从而在尽可能保持数据原貌的前提下最大限度精简数据量。

（5） 数据挖掘

在经过预处理的数据基础上，利用人工神经网络、遗传算法、决策树、规则推理等方法，高效地进行关联规则、序列模式、分类、聚集等各项分析。

（6） 结果的解释及评价

数据挖掘的目的在于根据最终用户的决策目的对提取的信息进行分析。上述过程将会得出一系列的分析结果、模式和模型。分析结果一般都是形式化的，这时需要通过可视化技术等技术手段，用图表、图形曲线等为用户提供清晰、直观的结果描述。在大多数情况下，对目标问题的描述是多侧面的，这时就要综合它们的规律性，进行进一步的抽象与过滤，提供合理的决策支持信息。

对结果进行评价也是一项必不可少工作。一种评价办法是直接使用原先建立的模型样本和样本数据进行检验，另一种办法是使用另一批数据，根据在这些数据上已知的规律性对其进行检验，再一种办法是在实际运行的环境中提取新鲜的数据进行检验。当评价的结果不能够令决策者满意时，需要重复以上的步骤，开始下一循环的知识发现过程。

4．数据挖掘技术和工具

数据发掘技术和工具可分为三大类：统计分析或数据分析，知识发现以及其他工具和技术，包括：可视化系统、地理信息系统、分形分析和私有工具。

（1） 统计分析

统计分析用于检查异常的数据模式，然后利用统计模型和数学模型解释这些数据模式。通常使用的模型有线性分析和非线性分析、连续回归分析和逻辑回归分析、单变量和多变量分析，以及时间序列分析。

统计分析工具用于一系列的商业活动中，寻求最佳机会来增加市场份额和利润，提高产品和服务的质量来使顾客更满意，通过流水线产品制造和后勤来增加利润。

直到最近，统计分析工具主要针对技术和工程应用中统计和技术上的专家。但是，许多企业面临缩小规模带来的人力和资金上的压力。因此，商业化统计分析工具，使其可以成功地为商业分析员所采纳并使用。这些商业分析员是领域内的专家，但却不是程序员或统计员。他们需要从数据仓库选择恰当数据，抽取它并进行分析。

（2） 知识发现

知识发现（Knowledge Discovery）源于人工智能和机器学习。通常人们将知识发现定义为：知识发现用一种简洁的方式从数据中抽取信息，这些信息是隐含的、未知的，并且是潜在有用的。或者，知识发现被看作是一种数据搜寻过程，它不必预先假设或提出问题，但仍能找到那些非预期的令人关注的信息，这些信息表示了数据元素的关系和模式，它也能通过完整的数据发现和数据分析找到商业规则。也有人简单地将知识发现看作，在数据仓库或数据集市几千兆字节的数据中找到预先未知的商业事实。

企业决策者和商业分析人员总是在寻找相关的和新的商业信息，以便做出更好的商业决策，这些决策对企业生命力有重要影响。使用传统的商业查询技术和数据分析技术时，要求所问的问题是恰当的。知识发现技术则由它自己来决定要问的问题，然后不断深入，作进一步探索，直到找到商业用户所寻求的知识。商业分析员没有时间，也没有那么多精力来从数据仓库中发现所有隐含的关系和模式。

知识发现是为了从数据仓库的大量数据中筛选信息，寻找经常出现的模式，检查趋势并发掘事实。知识发现系统试图让分析员作最少的指导，就可在最短的时间内找到事实和知识。因此在知识发现中，要查找数据仓库或数据集市的大量数据，找到事实或知识后，再将其发送给商业分析员。目前，商业分析人员主要利用商业知识和行业内的经验来区分有用的信息。这是人与计算机的理想结合。人脑具有在同一时间内分析多变量的最优算法，但其数据带宽有限。尽管计算机有巨大的带宽、精力和耐心，但它却无法理解多个商业变量。利用数据可视化工具和浏览工具有助于开发分析以前发掘的数据，以进一步增强信息发掘能力。

（3） 其他数据发掘技术和工具

可视化系统

可视化系统可给出带有多变量的图形化分析数据，帮助商业分析员进行发现，然后让商业分析员查看那些无论系统计算能力有多强，机器算法都极难确定的模式和关系。

有一种可视化技术表示的是平行坐标系，它可使商业分析员同时显示多个变量间的关系，它不像传统的可视化工具那样将坐标系映射为互相垂直的轴，而是映射为平行的轴。此种多变量数据集的描述保存了所有信息，并将多变量关系转化为良定义的二维模式，这有助于管理大量的数据，并有助于在复杂分析中应用可视化分析和查询方法。

地理信息系统

地理可视化系统中的不同物理位置直至地理表示都与仓库中的数据相关。商业分析员可以按地理环境来看待这些数据，并比较相同产品在不同地域的差异，或相同地域不同产品的差异。通过可视化一段时间内特定地理领域内销售的变化、产品售出服务等，也可以分析数据仓库中的临时数据。

分形分析

多维数据库提供了大量的分析信息并有较快的响应时间，但要存储整个数据仓库，则会受到空间限制。分形分析试图利用混沌科学来指明模式，然后用分行将其存储于数据仓库，其目的是要为大型数据库提供诸如 OLAP 风格的响应。

查询和报表工具

查询和报表工具（Query-and-Reporting Tools）与 QBE 工具、SQL 工具和典型数据库环境中的报表生成器类似。实际上，大部分数据仓库环境都支持诸如 QBE、SQL 和报表生成器之类的简单易用的数据处理子系统工具。数据仓库用户经常使用这类工具进行简单的查询并生成报表，如图 3-16 所示。

与 DBMS 一样，一个数据仓库系统具有一个搜索引擎。

图 3-16　数据挖掘工具集

智能代理

智能代理（Intelligent Agents）应用各种像神经网络、模糊逻辑这样的人工智能工具，形成 OLAP 中“信息发现”的基础。例如，华尔街某股票分析人员及应用一种叫做 Data/Logic 的 OLAP 软件，并加入神经网络为自己高效的股票和期货交易系统制定规则。还有一些 OLAP 工具与模糊逻辑结合来分析实时的技术处理过程。

智能代理代表了正在增长的各类进行信息处理的 IT 工具集中趋势。以前，智能代理被认为是人工智能领域的产物，很少被认为是一个企业中数据组织和管理部门的一部分。

小结

1．本章的主要内容：

a. 理解数据库在人工管理、文件管理、数据库系统的三个阶段的发展过程和特点
b. 数据库系统的体系结构
c. 理解数据库管理系统（DBMS）的功能及其工作过程
d. 了解多媒体数据库的组成
e. 了解数据仓库和数据挖掘的概念

2．本章的重点、难点：

数据库在人工管理、文件管理、数据库系统的三个阶段的发展过程

习题

选择题

1．共享和结构化是（　）管理数据的特点
A. 人工　B. 文件系统　C. 数据库系统　D. 以上三者
2．数据库体系结构的内模式又称为（　）
A. 子模式　B. 概念模式　C. 存储模式　D. 用户模式
3．实体关系图中，实体之间存在（　）的关系
A. 1 对 1　B. 1 对多　C. 多对多　D. 以上三者
4．医生实体与患者实体之间存在（　）的关系
A. 1 对 1　B. 1 对多　C. 多对多　D. 以上三者
5．部门实体与员工实体之间存在（　）的关系
A. 多对 1　B. 1 对多　C. 多对多　D. 1 对 1
6．部门实体与员工实体之间，部门实体处于（　）方
A. 1　B. 多　C. 不一定　D. 以上三者
7．部门实体与员工实体之间，员工实体处于（　）方
A. 1　B. 多　C. 不一定　D. 以上三者
8．采用树形结构组织数据的数据库模型是（　）
A. 层次模型　B. 网状模型　C. 关系模型　D. 不一定
9．采用二维表结构组织数据的数据库模型是（　）
A. 层次模型　B. 网状模型　C. 关系模型　D. 不一定
10．采用（　）可以实现对数据的插入、删除和修改等操作
A. 数据描述语言　B. 数据操纵语言　C. 都可以　D. 都不行

思考题

了解数据库发展的三个基本过程，数据库的三级模式；了解多媒体数据库和数据仓库的基本概念。

1. 计算机在数据处理各个阶段有哪些基本特点？
2. 什么是数据库的三级模式？
3. 为什么说数据库系统具有程序和数据的独立性。

填空题

1. E-R 图可以描述实体之间_____、_____和_____关系。
2. 数据库的三级模式包括_____、_____和_____。

3. 关系是_____表，是可以保存在计算机中的一个_____。
4. 关系的传统集合运算包括_____、_____、等。
5. 数据是指存储在某一媒体上可加以鉴别的_____，媒体可以包括_____、_____、_____、_____等等种类。
6. 信息是来自于现实世界_____或运动形态的集合，是人们进行各种活动所需要的_____。