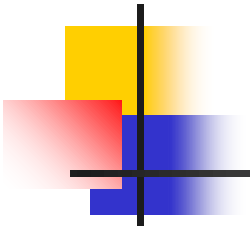
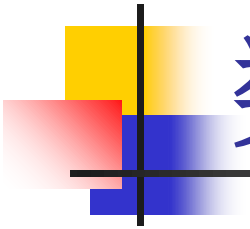


数据仓库与数据挖掘概述



数据仓库与数据挖掘概述

1. 数据仓库的兴起
2. 数据挖掘的兴起
3. 数据仓库和数据挖掘的结合

1. 从数据库到数据仓库

（1）“数据太多，信息不足”的现状

（2）异构环境的数据的转换和共享

（3）利用数据进行数据处理转换为利用数据支持决策

1.1 数据库用于事务处理

- 数据库中存放的数据基本上是保存当前数据，随着业务的变化随时在更新数据库中的数据。
- 不同的管理业务需要建立不同的数据库。例如，银行中储蓄业务、信用卡业务分别要建立储蓄数据库和信用卡数据库。
- 数据库是为满足事务处理需求建立的，在帮助人们进行决策分析时显得不适用。

1.2 数据库系统的局限性

数据库系统作为数据管理手段，主要用于事务处理，取得了巨大的成功，那么能否将它应用于分析型数据处理呢？答案是否定的，主要原因包括以下几点。

（1）数据的分散

联机事务处理的目的在于使业务处理自动化，一般只需要与本部门业务有关的当前数据，而对整个企业范围内的集成应用考虑很少。企业内部事务处理的应用之间实际上几乎都是独立的，造成了当前绝大部份企业内数据的真正状况是分散而非集成的。

出现这种现象有多重原因。有设计方面的、有经济方面的、还有体制方面，以及历史、地理方面等。

1.2 数据库系统的局限性

（2）“蜘蛛网”问题

解决数据分散的一种方法是对数据进行集成。在联机事务处理系统出现不久，就出现一种称作“抽取”处理的程序”。用户利用抽取程序从各个分散的数据库中查找有用的数据。然后这些数据被提取出来放入新的文件或数据库中，供用户使用。由于抽取程序能将数据从联机事务处理系统转移出来，对这些数据进行分析时不会影响联机事务处理系统的效率，因此，受到程序员的喜爱，被大量应用。

这些经抽取得到的新文件或数据库又被某些用户再进行抽取，这种不加以控制的连续抽取最终导致系统内数据间形成了错综复杂的网状结构，人们形象地称为”蜘蛛网”。企业的规模越大，”蜘蛛网”问题就越严重。

1.2 数据库系统的局限性

（3）数据不一致问题；

由于前述的数据分散和“蜘蛛网”等问题，导致了多个应用间的数据不一致。具体表现有：

（1）同一字段在不同的应用中具有不同的数据类型。

（2）同一字段在不同的应用中具有不同的名字。

（3）同名字段，不同含义。

为了将这些不一致的数据集成起来，首先必须对它门进行转换，消除不一致之后才能供分析使用。因此，这是一项很繁重的工作。

1.2 数据库系统的局限性

（4）数据动态集成问题

由于每次分析都进行数据集成的开销很大，一些应用仅在开始对所需数据进行了集成，以后就一直以这部分集成的数据作为分析的基础，不再与数据发生联系，我们称这种方式的集成为静态集成。静态集成的最大缺点在于，如果在数据集成后数据源中数据发生改变，这些变化不能反映给决策者，导致决策者使用的是过时的数据。

集成数据必须以一定的周期进行刷新，我们称其为动态集成。显然联机事务处理系统不具备动态集成的能力。

1.2 数据库系统的局限性

（5）历史数据问题

联机事务处理一般只需要当前数据，在数据库中一般也只存储短期内的数据，且不同数据的保存期限也不同。即使被保留的历史数据，也没有得到充分利用。

对于决策分析而言，历史数据是相当重要的，许多分析方法必须以大量的历史数据为依据，没有对历史数据的详细分析，是难以把握企业的发展趋势的。

1.2 数据库系统的局限性

（6）数据的综合问题

在联机事务处理系统中积累了大量的细节数据。一般而言，DSS并不对这些细节数据进行分析。这主要有两个原因，一是细节数据量太大，会严重影响分析的效率；二是太多的细节数据不利于分析人员将注意力集中于有用的信息上。因此，在分析前，往往需要对细节数据进行不同程度的综合。而事务处理系统不具备这种能力，根据规范化理论，这种综合还往往因为是一种数据冗余而加以限制。

1.2 数据库系统的局限性

小结：

传统数据库所能做到的只是对已有的数据进行存取以及简单的查询统计，即使是一些流行的OLAP工具，也无非是另一种数据展示方式而已。人们仍然无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。这也直接导致了目前“数据爆炸但知识匮乏”的现状。

1.3 数据仓库用于决策分析

- 数据库用于事务处理，数据仓库用于决策分析
- 数据库保持事务处理的当前状态，数据仓库既保存过去的数据又保存当前的数据
- 数据仓库的数据是大量数据库的集成
- 对数据库的操作比较明确，操作数据量少。对数据仓库操作不明确，操作数据量大

1. 4 数据库与数据仓库对比

数据库	数据仓库
细节的	综合或提炼的
在存取时准确的	代表过去的数据
可更新的	不更新
一次操作数据量小	一次操作数据量大
面向应用	面向分析
支持管理	支持决策

1. 5 数据仓库与数据库的关系

- **数据仓库与数据库的关系**

- 数据库的应用包括：事务型应用和分析型应用
- 物理数据库实际存储的数据包括：

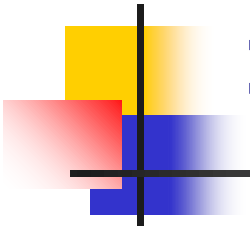
  事务型数据（或称操作数据）和分析型数据（也可称为汇总数据、信息数据）。
- 起初，两类数据放到一起，即分散存储在各底层的业务数据库中。
- 后来，随着企业规模的扩展、数据量的增加、以及希望在决策分析时得到更多支持需求的日益迫切，并且考虑保证原有事务数据库的高效性与安全性。因此将分析型数据与事务型数据相分离，单独存放，即形成了所谓的数据仓库。

1. 5 数据仓库与数据库的关系

数据仓库只不过是因为用户需求增加而对某一类数据库应用的一个范围的界定。单就其是数据的存储容器这一点而言，数据仓库与数据库并没有本质的区别。

而且在更多的时候，我们是将数据仓库作为一个数据库应用系统来看待的。

因此，不应该说数据库到数据仓库是技术的进步。



1.6 从OLTP到OLAP

1.联机事物处理（OLTP）

2.联机分析处理（OLAP）

3.OLTP与OLAP的对比

1.6.1联机事物处理（OLTP）

- 联机事物处理（**On Line Transaction Processing，OLTP**）是在网络环境下的事务处理工作，以快速的响应和频繁的数据修改为特征，使用户利用数据库能够快速地处理具体的业务。
- **OLTP**是用户的数据可以立即传送到计算中心进行处理，并在很短的时间内给出处理结果。也称为实时系统**(Real time System)**。**OLTP**主要用于包括银行业、航空、邮购订单、超级市场和制造业等的输入数据和取回交易数据。如银行为分布在各地的自动取款机 **(ATM)**完成即时取款交易；机票预定系统能每秒处理的定票事务峰值可以达到**20000**个。

1.6.1 联机事物处理（OLTP）

- **OLTP**的特点在于事务处理量大，应用要求多个并行处理，事务处理内容比较简单且重复率高。
- 大量的数据操作主要涉及的是一些增加、删除、修改、查询等操作。每次操作的数据量不大且多为当前的数据。
- **OLTP**处理的数据是高度结构化的，数据访问路径是已知的，至少是固定的。
- **OLTP**面对的是事务处理操作人员和低层管理人员。
- 但是，为高层领导者提供决策分析时， **OLTP**则显得力不从心。

1.6.2 联机分析处理（OLAP）

- **E.F.Codd**认为决策分析需要对多个关系数据库共同进行大量的综合计算才能得到结果。
- **E.F.Codd**在**1993**年提出了多维数据库和多维分析的概念，即联机分析处理（**OnLine Analytical Processing，OLAP**）概念。
- 关系数据库是二维数据（平面），多维数据库是空间立体数据。

新的挑战：如何不被淹没在信息的海洋里

1.6.2 联机分析处理（OLAP）

- **OLAP**专门用于支持复杂的决策分析操作，侧重对分析人员和高层管理人员的决策支持，
- **OLAP**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地形式将查询结果提供给决策制定人。
- **OLAP**软件，以它先进地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。
- **OLAP**的基本思想是决策者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。

1.6.3 OLTP与OLAP的对比

OLTP	OLAP
细节性数据	综合性数据
当前数据	历史数据
经常更新	不更新，但周期性刷新
一次性处理的数据量小	一次处理的数据量大
对响应时间要求高	响应时间合理
面向应用，事务驱动	面向分析，分析驱动

1.7 数据仓库的定义与特点

1. 数据仓库定义

（1）W. H. Inmon在《建立数据仓库》一书中，对数据仓库的定义为：

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决策制定过程。

（2）SAS软件研究所观点：

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。

1.7 数据仓库的定义与特点

“数据仓库之父”W.H.Inmon在其《Building the Data Warehouse》一书中，指出数据仓库中的数据应具备以下4个基本特征：

（1）数据仓库的数据是面向主题的；

（2）数据仓库的数据是集成的；

（3）数据仓库的数据是不可更新的；

（4）数据仓库的数据是随时间不断变化的。

并且给出了数据仓库的定义：数据仓库是一个面向主题的、集成的、不可更新的、随时间不断变化的数据集合，用以更好地支持企业或组织的决策分析处理。

1.7 数据仓库的定义与特点

（1）数据仓库是面向主题的

是相对于传统数据库的面向应用而言的。所谓面向应用，指的是系统实现过程中主要围绕着一些应用或功能。而面向主题则考虑一个个的问题域，对问题域涉及到的数据和分析数据所采用的功能给予同样的重视。**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，银行的数据仓库的主题：客户。DW的客户数据来源：从银行储蓄DB、信用卡DB、贷款DB等三个DB中抽取同一客户的数据整理而成。在DW中能全面地分析客户数据，再决定是否继续给予贷款。

1.7 数据仓库的定义与特点

（2）数据仓库是集成的

最重要的特点。数据仓库中的数据来自各个不同的数据源（操作数据库）。由于历史的原因，各操作数据库的组织结构往往是不同的，在这些异构数据输入到数据仓库之前，必须经历一个集成过程。

对不同的数据来源进行统一数据结构和编码。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。

将原始数据结构做一个从面向应用到面向主题的大转变。

1.7 数据仓库的定义与特点

（3）数据仓库是稳定的(不可修改的)

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DSS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。

1.7 数据仓库的定义与特点

（5）数据仓库的数据量很大

大型DW的数据是一个TB（1000GB）级数据量（一般为10GB级DW，相当于一般数据库100MB的100倍）

（6）数据仓库软、硬件要求较高

需要一个巨大的硬件平台

需要一个并行的数据库系统

1.8 数据仓库的体系结构

1.8.1 体系结构

图1 是一个典型的数据仓库系统的体系结构图。数据仓库系统由数据源、集成工具、数据仓库与数据仓库服务器、OLAP服务器、元数据与元数据管理工具、数据集市和前台分析工具等组成。

1.8 数据仓库的体系结构

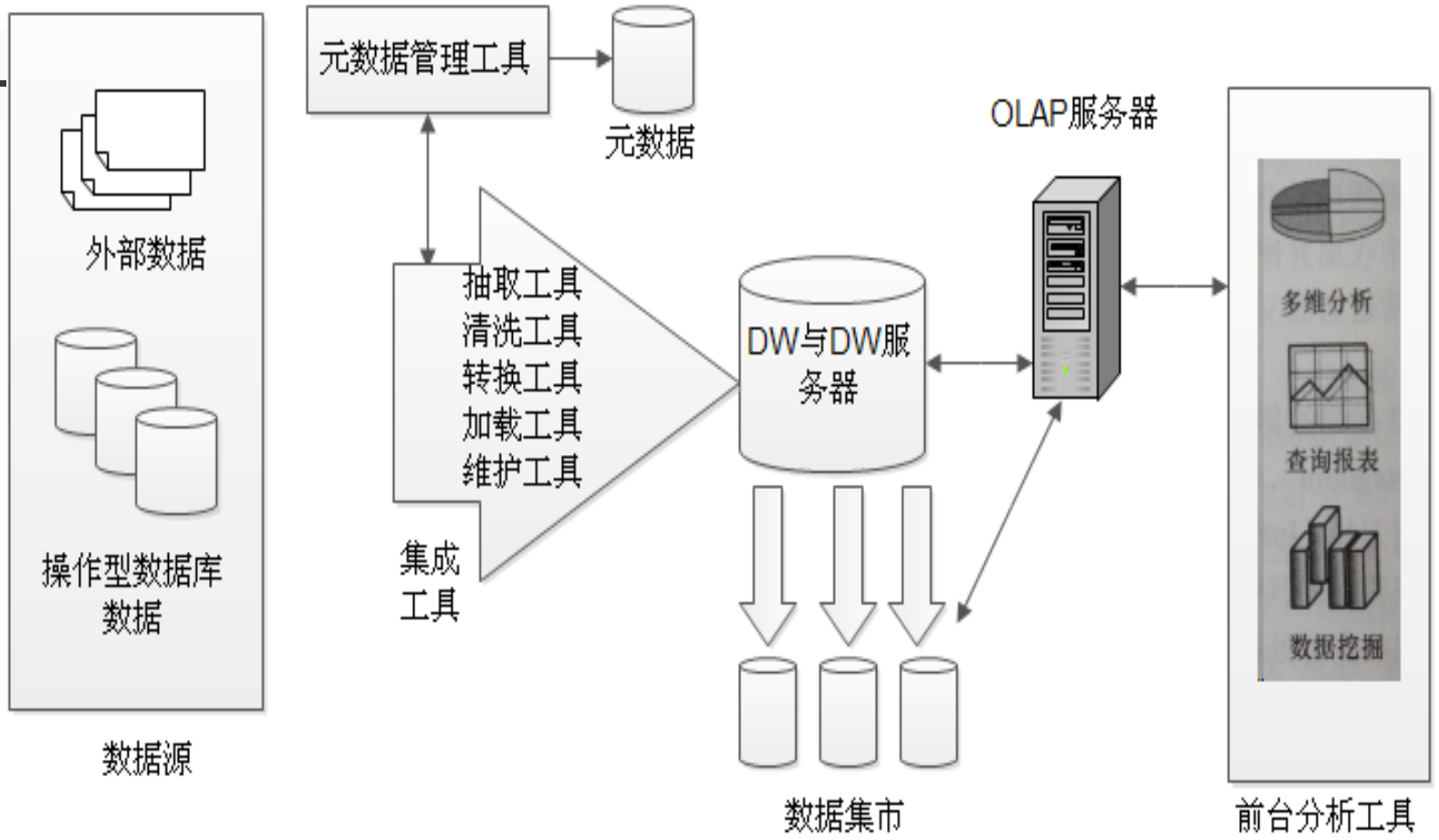


图1 典型的数据仓库系统的体系结构

1.8 数据仓库的体系结构

数据源是数据仓库系统的基础，是整个系统的数据源泉。通常包括企业内部信息和外部信息。内部信息包括存放于企业操作型数据库中的各种业务数据等；外部数据包括各类法律法规、市场信息、竞争对手信息及各类外部统计数据及各类文档等。

集成工具包括数据抽取（Extracting）、清洗（Cleaning）、转换（Transformtion）、加载（Load）工具，简称为ETL工具，完成数据的集成。

数据抽取，就是从数据源中选择数据仓库需要的数据。数据抽取的技术难点在于要针对不同平台、不同结构、不同厂商的数据库，设计不同的抽取工具。

1.8 数据仓库的体系结构

数据清洗，为保证数据质量，对抽取到的数据要进行清洗，例如，消除不一致、统一计量单位、确认默认值。

数据转换，是将清洗后的数据按照数据仓库的主题进行组织。

数据加载，就是将数据装入数据仓库中。

此外，ETL还负责建立元数据。元数据主要说明数据仓库中数据的来源、加工过程等。

数据仓库服务器，负责管理数据仓库中的数据，存储企业级的数据，为整个企业的数据分析提供一个完整的、统一的视图。一般由关系数据库管理系统扩展而成。

1.8 数据仓库的体系结构

OLAP服务器，对分析需要的数据按照多维数据模型进行再次重组，以支持用户多角度、多层次的数据分析。其具体实现可以分为：ROLAP、MOLAP、HOLAP以及特殊的SQL服务器。

（1）ROLAP(Reletion OLAP)结构：采用关系DBMS或扩展的关系DBMS来存储和管理数据，用OLAP服务器提供聚集计算、查询优化等功能。

（2）MOLAP(Multi-dimension OLAP)结构：采用多维数组来存储数据。多维数组存储的优点是可以对数据进行直接定位、计算速度快、不需要索引，但在数据稀疏的情况下需进行数据压缩，以减少存储空间。

1.8 数据仓库的体系结构

（3）HOLAP(Hybrid OLAP)结构：该结构将ROLAP和MOLAP结合起来，既利用了ROLAP可扩展性好的优点，也利用了MOLAP计算速度快的优点。例如，可以将细节数据存放在关系数据库中，而将综合数据存储在MOLAP服务器中。微软的SQL server 7.0采用的就是这种结构。

（4）特殊SQL服务器：为满足日益增长的OLAP处理需要，一些关系数据库或数据仓库厂商对原有系统的扩展，为只读环境下在星型模型或雪片模型基础上进行SQL查询提供支持。

1.8 数据仓库的体系结构

数据集市，是一种小型数据仓库。它通常有较少的主题域，因此细节数据和历史数据较少，是部门级的。是面向部门级的应用，也称之为部门级数据仓库。

前台分析工具，主要包括各种数据分析工具，如报表工具、数据挖掘工具等。

元数据，是整个数据仓库的所有描述性信息。

1.8 数据仓库的体系结构

从图1 可以看出，在数据仓库系统中，数据从数据源到最终的分析结果呈现给用户，中间需要经过一系列的过程。

（1）抽取适当的数据源。只选取对现在和将来决策和分析有用的业务数据进行抽取就可以了。

（2）转换、清洗、重构等数据加工过程。

（3）建立海量、高效的企业级数据仓库。这个企业级数据仓库必须能够在海量数据基础上服务于大量并发用户，且无论是数据处理速度还是查询速度都应该满足一定的要求。

1.8 数据仓库的体系结构

（4）针对特定的分析主题，建立专门的数据集市。

（5）针对特定的业务问题，使用特殊的数理统计算法进行数据挖掘。

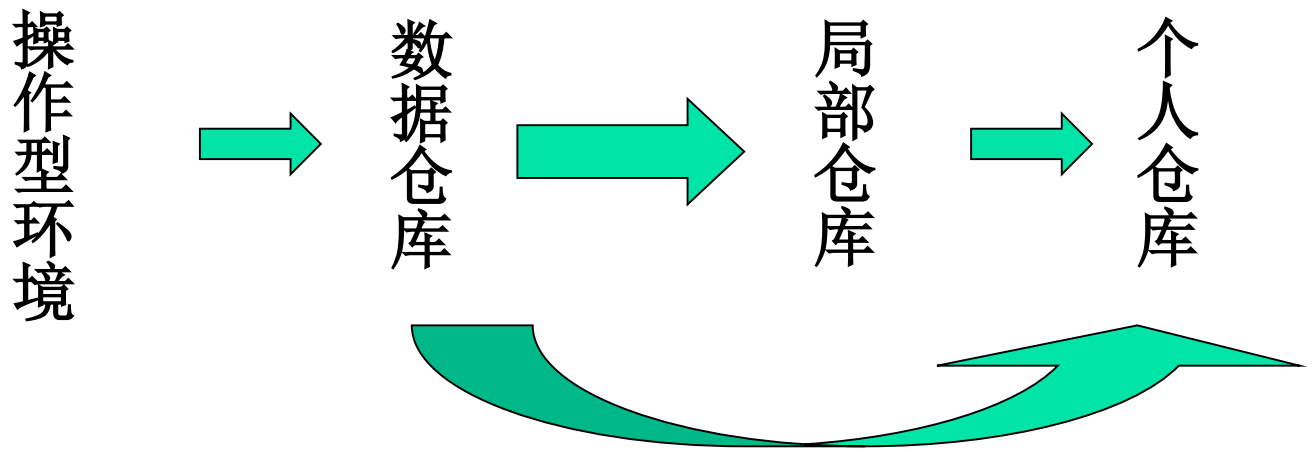
（6）前端展现应用。最终用户的界面必须简单易用且功能强大，必须具有良好的权限控制，必须易于维护。

1.9 数据库体系化环境

数据库体系化环境是在一个企业或组织内，由各面向应用的OLTP数据库及各级面向主题的数据仓库所组成的完整的数据环境，在这个数据环境上建立和一个企业或部门的从联机事务处理到企业管理决策的所有应用。

这个数据环境分为两个部分：操作型环境和分析型环境，分别为操作型处理和分析型处理这两种不同的数据处理服务。

1.9 数据库体系化环境



上图所示是对体系化环境的一个简单描述，它分为四个层次：操作型环境，全局级数据仓库，部门级的局部仓库和个人的数据仓库。其中部门级的数据仓库也称为数据集市。

1.9 数据库体系化环境

数据从操作型环境经过综合整理进入全局数据仓库，企业中的有关部门再从全局数据仓库中组织适合自己特殊分析需求的数据，建立自己的局部仓库；而个人不仅可以从全局数据仓库中提取数据，而且可以从部门级局部仓库中提取所需数据。这样，由于数据在全局仓库中都已经是集成的，一致的，所以部门和个人的抽取工作效率将会很高，并且不会出现之前提到的“蜘蛛网”情况。

1.9 数据库体系化环境

四层体系化环境可以很好地与企业实际的部门组织结构对应起来，在实际中，管理工作可分为高层，中层及基层三种。基层管理也称为操作管理，其主要任务是一般日常业务处理。操作型环境就是面向这一层次，如进行联机事务处理。中层管理既包含一般业务处理，又需进行简单分析，作出一般的决策和控制。部门级的局部数据仓库面向这一层。高层管理的主要任务是进行战略决策，需要进行复杂的分析加工，个人级数据仓库可以说面向这一层次。

1.10 数据集市

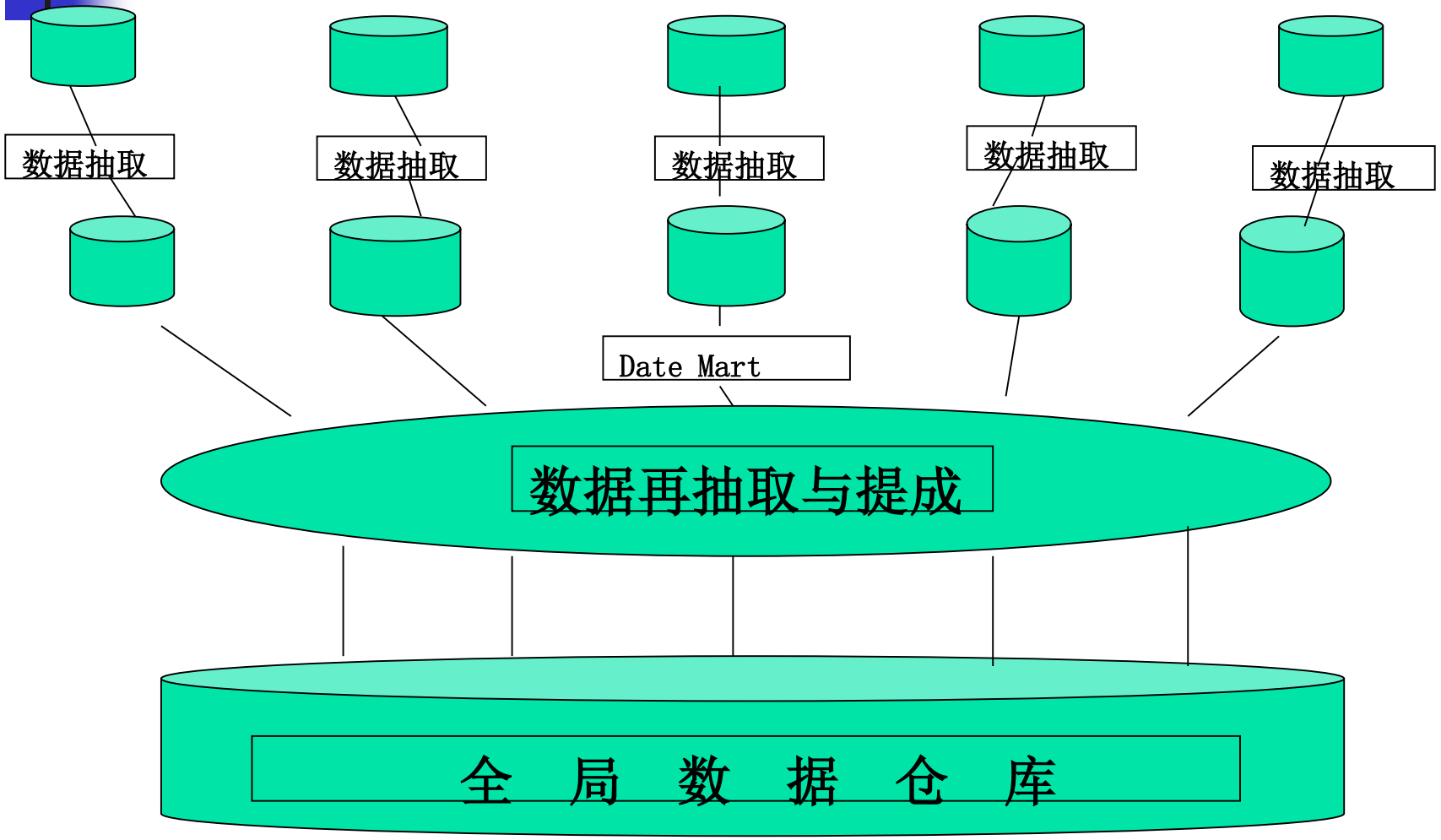
在四级体系化环境中，提出了数据仓库的几个不同的层次或称为不同的级别：全局级数据仓库，部门级数据仓库和个人的数据仓库。从图中的数据抽取关系来看，这三层数据仓库的实现应该采用“自顶向下”的方法，也就是说，先建立一个全局的数据仓库结构，然后再在这一全局数据仓库的基础上建立部门级和个人级的数据仓库，这样的建设途径有利于各级数据仓库的一致性的控制。

1.10 数据集市

但是，“自顶向下”地建设多级数据仓库必然要求先从原来分散的操作环境中来建立一个全局数据仓库。而全局级数据仓库的规模往往较大，在原来分散的操作型环境基础上建立这第一个大而全的数据仓库，其实施周期过长，见效慢，费用昂贵，这些往往是许多企业不愿意或不能承担的。

数据集市的组织标准是多样的，除了上述的按业务来划分外，也可以按照数据仓库的主题或数据的地理分布来组织。

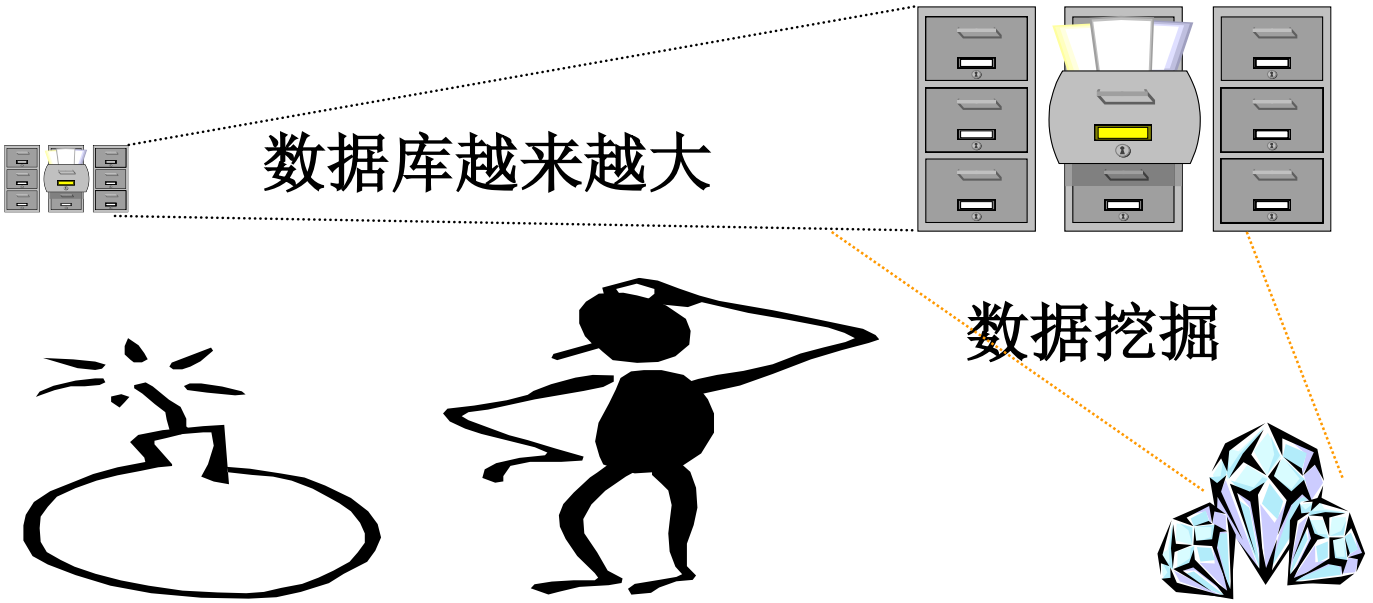
1.10 数据集市



2 数据挖掘的兴起

二十世纪末以来，全球信息量以惊人的速度急剧增长—据估计，每二十个月将增加一倍。许多组织机构的IT系统中都收集了大量的数据（信息）。目前的数据库系统虽然可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。为了充分利用现有信息资源，从海量数据中找出隐藏的知识，数据挖掘技术应运而生并显示出强大的生命力。

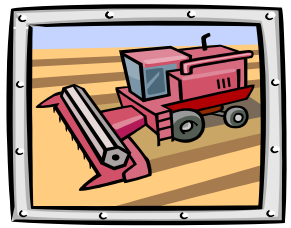
Why?数据挖掘的社会需求



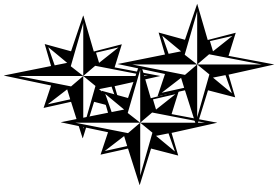
2 数据挖掘的兴起

所有企业面临的一个共同问题是：企业数据量非常大，而其中真正有价值的信息却很少，因此需要从大量的数据中经过深层分析，获得有利于商业运作、提高竞争力的信息，就像从矿石中淘金一样，数据挖掘也由此而得名。

矿山（数据）



挖掘工具（算法）



金子（知识）

概述

数据挖掘是八十年代投资AI研究项目失败后，AI转入实际应用时提出的。它是一个新兴的，面向商业应用的AI研究。

1989年8月，在美国底特律召开的第11届国际人工智能联合会议的专题讨论会上首次出现数据库中的知识发现（Knowledge Discovery in Database，KDD）这一术语。

随后，在1991年、1993年和1994年都举行KDD专题讨论会，汇集来自各个领域的研究人员和应用开发者，集中讨论数据统计、海量数据分析算法、知识表示、知识运用等问题。最初，数据挖掘是作为KDD中利用算法处理数据的一个步骤，其后逐渐演变成KDD的同义词。

概述

(1) 1980年在美国召开了第一届国际机器学习研讨会
明确了机器学习是人工智能的重要研究方向
(2) 1989年8月于美国底特律市召开的第一届知识发现（KDD）国 际学术会议；
首次提出知识发现概念
(3) 1995年在加拿大召开了第一届知识发现和数据挖掘（DM）国际学术会议；
首次提出数据挖掘概念

IEEE的Knowledge and Data Engineering会刊率先在1993年出版了KDD技术专刊。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论。数据挖掘已经成了国际学术研究的重要热点之一。

数据挖掘定义

- 技术角度的定义

数据挖掘（Data Mining)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、人们事先不知道的、但又是潜在有用的信息和知识的过程。与数据挖掘相近的同义词包括：数据融合、数据分析和决策支持等。

这一定义包括好几层含义:数据源必须是真实的、海量的、含噪声的;发现的是用户感兴趣的知识;发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

商业角度的定义

数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性信息。

简言之，数据挖掘其实是一类深层次的数据分析方法。因此，数据挖掘可以描述为:按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的有效方法。

数据挖掘与传统分析方法的区别

一种深层次的数据分析方法。

数据分析本身已有多年的历史，只不过在过去数据收集和分析的一般目的是用于科学研究；另外，由于当时计算能力的限制，很难实现大量数据的复杂分析。

现在，由于各行业业务自动化的实现，商业领域产生了大量的业务数据，这些数据并不是为了分析的目的而收集的，而是在商业运作过程中由于业务需要而自然产生的。数据挖掘不再是单纯为了研究，更主要的是为商业决策提供真正有价值的信息，进而获得利润。

数据挖掘与传统数据分析方法区别

（1）数据挖掘的数据源与以前相比有了显著的改变；

数据挖掘出现的背景是“数据爆炸但知识贫乏”，数据是海量的；数据有噪声；数据可能是非结构化的；

（2）传统的数据分析方法一般都是先给出一个假设然后通过数据验证，在一定意义上是假设驱动的；与之相反，数据挖掘在一定意义上是发现驱动的，模式都是通过大量的搜索工作从数据中自动提取出来 。即数据挖掘是要发现那些不能靠直觉发现的信息或知识，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可能越有价值。在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了小孩尿布和啤酒之间有着惊人的联系。

数据挖掘与其他科学的关系

数据挖掘作为一门新兴的交叉学科，涉及数据库系统、数据仓库、统计学、机器学习、可视化、信息检索和高性能计算等诸多领域。

此外，还与神经网络、模式识别、空间数据分析、图像处理、信号处理、概率论、图论和归纳逻辑等等领域关系密切。

国内研究

与国外相比，国内对数据挖掘的研究起步稍晚，但发展势头强劲。我国于1987年召开了第一届全国机器学习研讨会。

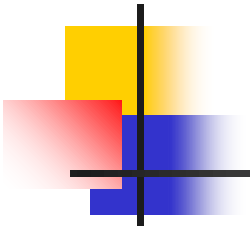
1993年，国家自然科学基金首次资助复旦大学对该领域的研究项目。

目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究。

发展趋势

近年来，数据挖掘的研究重点逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多学科之间的相互渗透。

例如，1998年在美国纽约举行的第四届知识发现与数据挖掘国际学术会议不仅进行了学术讨论，并且有30多家软件公司展示了他们的数据挖掘软件产品，不少软件已在北美、欧洲等国得到应用。



展望

- 未来的热点应用领域

网站的数据挖掘（Web site data mining）

生物信息或基因的数据挖掘

文本挖掘（Textual mining）

多媒体挖掘

网站的数据挖掘（Web site data mining）

当前Internet上各类电子商务网站风起云涌，电子商务业务的竞争比传统的业务竞争更加激烈。客户从一个电子商务网站转换到竞争对手那边，只需点击几下鼠标即可，电子商务环境下客户保持比传统商业更加困难。若想在竞争中生存进而获胜，您必须比竞争对手更了解客户。电子商务网站每天都可能有上百万次的在线交易，生成大量的记录文件（Log files）和登记表，如何对这些数据进行分析和挖掘，及时地了解客户的喜好、购买模式，甚至是客户一时的冲动，设计出满足于不同客户群体需要的个性化网站，进而增加竞争力，几乎变得势在必行。

生物信息或基因的挖掘

生物信息或基因数据挖掘则完全属于另外一个领域，在商业上很难讲有多大的价值，但对于人类却受益非浅。例如，基因的组合千变万化，得某种病的人的基因和正常人的基因到底差别多大？能否找出其中不同的地方，进而对其不同之处加以改变，使之成为正常基因？这都需要数据挖掘技术的支持。

对于生物信息或基因的数据挖掘和通常的数据挖掘相比，无论在数据的复杂程度、数据量还有分析和建立模型的算法方面，都要复杂得多。从分析算法上讲，更需要一些新的和高效的算法。现在很多厂商正在致力于这方面的研究。但就技术和软件而言，还远没有达到成熟的地步。

文本挖掘（Textual mining）

文本挖掘是人们关心的另外一个话题。例如，在客户服务中心，把同客户的谈话转化为文本数据，再对这些数据进行挖掘，进而了解客户对服务的满意程度和客户的需求以及客户之间的相互关系等信息。

无论是在数据结构还是在分析处理方法方面，文本数据挖掘和数据挖掘相差很大。文本挖掘并不是一件容易的事情，尤其是在分析方法方面，还有很多需要研究的专题。目前市场上有一些类似的软件，但大部分方法只是把文本移来移去，或简单地计算一下某些词汇的出现频率，并没有真正实现语义上的分析功能。

多媒体挖掘（Multimeadia Mining）

（1）基于描述的检索系统

- 基于图像的描述创建索引并实现对象检索，如关键字、标题、尺寸和创建时间等；
- 人工实现则极为费时、费力；
- 自动实现则往往结果不理想。

（2）基于内容的检索系统

- 支持基于图像内容的检索，例如颜色、质地、形状、对象及小波变换

数据挖掘与OLAP的比较

（1） **OLAP**的多维分析

在带层次的纬度和跨纬度进行多维数据分析。功能包括聚合、分配、比率、乘积等描述性的建模功能。

OLAP的典型应用，通过商业活动变化的查询发现的问题，经过追踪查询找出问题出现的原因，达到辅助决策的作用。

（2）数据挖掘

数据挖掘是以变量和记录为基础进行分类。任务在于聚类（如神经网络聚类）、分类（如决策树分类）、预测等，带有探索性的建模功能。

3 数据仓库和数据挖掘的结合

3.1 数据仓库和数据挖掘的区别与联系

3.2 基于数据仓库的决策支持系统

3.1 数据仓库与数据挖掘的区别

- 数据仓库是一种存储技术，将大量数据按决策需求进行重新组织，为用户提供辅助决策的随机查询、综合信息以及随时间变化的趋势分析。它能适应于不同用户对不同决策需要提供所需的数据和信息。

- 数据挖掘是从机器学习人工智能发展起来的。研究各种方法和技术，从大量的数据中挖掘出有用的信息和知识。

3.1 数据仓库与数据挖掘的区别

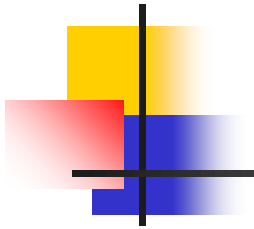
- 数据仓库与数据挖掘都是决策支持新技术。但它们有着完全不同的辅助决策方式。
- 在数据仓库系统的前端的分析工具中，数据挖掘是其中重要工具之一。它可以帮助决策用户挖掘数据仓库的数据中隐含的规律性。数据挖掘用于数据仓库实现决策支持：

（1）预测客户购买倾向；
（2）客户利润贡献度分析；
（3）分析欺诈行为；
（4）销售渠道优化分析等。

数据仓库和数据挖掘的结合对支持决策会起更大的作用。

3.2 基于数据仓库的决策支持系统

- 数据仓库中有大量的综合数据，为决策者提供了综合信息。数据仓库保存有大量历史数据，通过预测模型计算可以得到预测信息。
- 联机分析处理（**OLAP**）对数据仓库中的数据进行多维数据分析，即多维数据的切片、切块、旋转、钻取等，得到更深层中的信息和知识。
- 数据挖掘（**DM**）技术能获取关联知识、时序知识、聚类知识、分类知识等。
- 数据仓库（**DW**）、联机分析处理（**OLAP**）、数据挖掘（**DM**）等结合，形成决策支持系统。



结　束

End