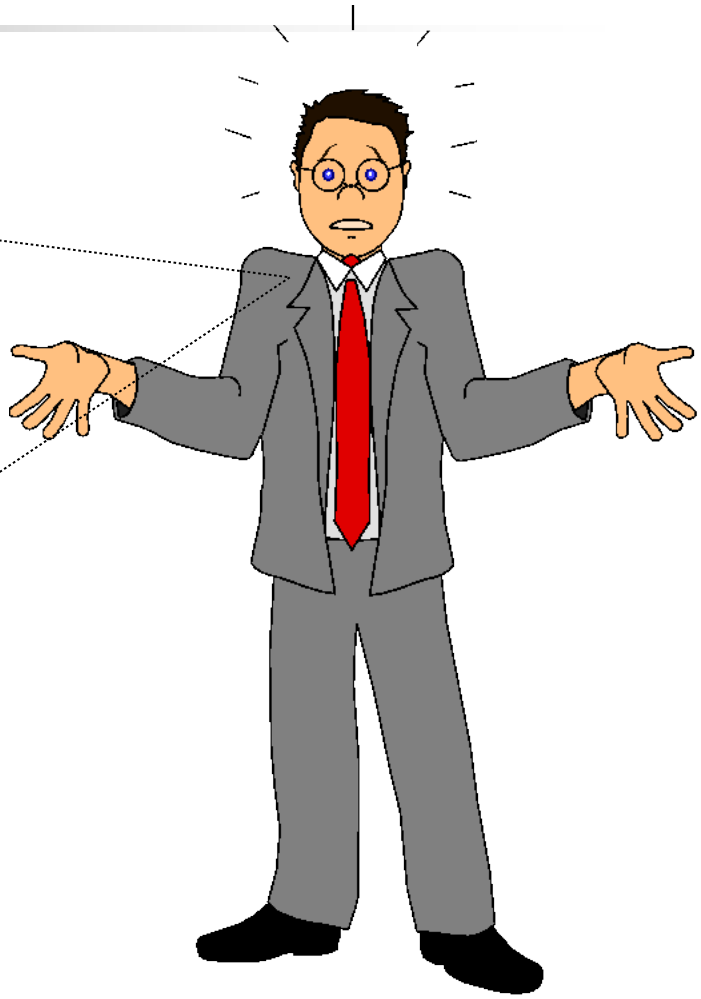
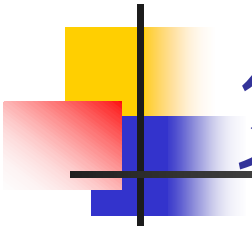
数据仓库与数据挖掘

任课教师：
工作单位：
办公地点：
联系电话：
QQ号码：



第1章

数据仓库与数据挖掘概述



第1章

1.1 数据仓库的兴起

1.2 数据挖掘的兴起

1.3 数据仓库和数据挖掘的结合

1.1.1 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.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。（举例）

- **数据库的局限性**

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。

2.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

3.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

➢ 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。

- **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。

1.1.2从OLTP到OLAP

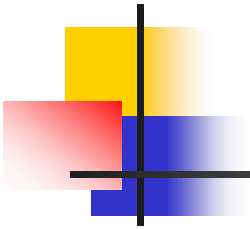
1.联机事物处理（OLTP）

2.联机分析处理（OLAP）

3.OLTP与OLAP的对比

1.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。

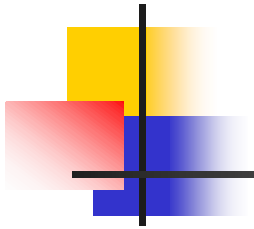


- OLTP的特点在于事务处理量大，应用要求多个并行处理，事务处理内容比较简单且重复率高。
- 大量的数据操作主要涉及的是一些增加、删除、修改、查询等操作。每次操作的数据量不大且多为当前的数据。
- OLTP处理的数据是高度结构化的，数据访问路径是已知的，至少是固定的。
- OLTP面对的是事务处理操作人员和低层管理人员。
- 但是，为高层领导者提供决策分析时， OLTP则显得力不从心。

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（On Line 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



- **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。

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

1.1.4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。

2. 数据仓库特点

（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。**

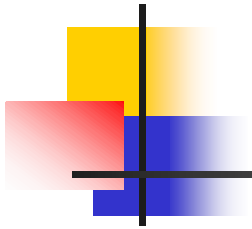
例如，银行的数据仓库的主题：客户。DW的客户数据来源：从银行储蓄DB、信用卡DB、贷款DB等三个DB中抽取同一客户的数据整理而成。在DW中能全面地分析客户数据，再决定是否继续给予贷款。

（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。



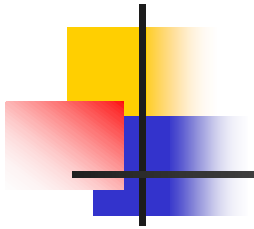
（3）数据仓库是稳定的(不可修改的)

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。



（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

1.2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