数据仓库与联机分析处理技术

北京大学信息科学与技术学院

童云海

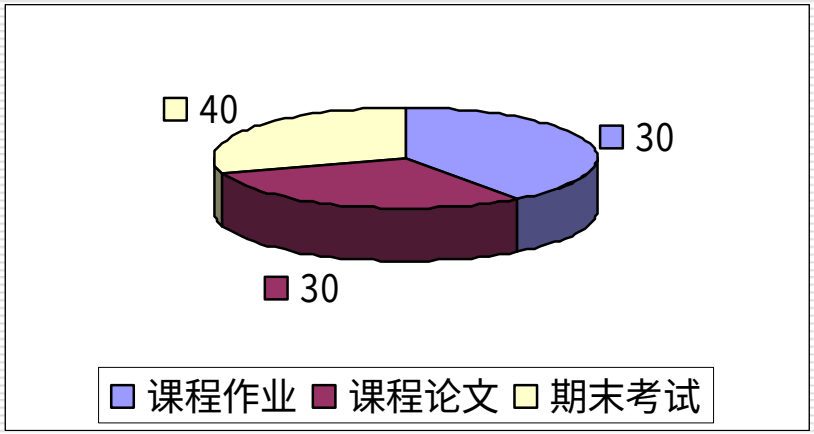
2005年3月

关于本课程

- 讲课与实习相结合的方式
- 要求
  - 强调掌握（理解）内容的本质，不主张死记硬背
  - 强调实践，不主张仅应付书面的考试

课程考核办法

- 课程作业
  - 2－3次作业
  - 比例：30%
- 课程论文
  - 比例：40%
- 期末考试
  - 比例：30%



参考书目

- Building the Data Warehouse(3rd Edition) W.H.Inmon 2003.3
- The Data Warehouse Toolkit(2nd Edition) R.Kimball 2002
- 数据仓库/Building the Data Warehouse(3rd Edition) W.H.Inmon 著，黄厚宽 等译 机械工业出版社 2003.3
- 王珊，《数据仓库与联机分析处理技术》，科学出版社，1998年

课程主要内容

- 第一章 数据仓库基本概念
- 第二章 联机分析处理的概念
- 第三章 数据仓库的设计
- 第四章 数据仓库的实现
- 第五章 ETL的设计和实现
- 第六章 OLAP Server的相关技术
- 第七章 操作数据存储（ODS）
- 第八章 数据仓库系统
- 第九章 数据仓库技术的发展现状

课程学习目标

- 理解并掌握数据仓库与联机分析处理的基本概念
- 获得数据仓库的构建和应用的一些方法
- 掌握数据仓库建模的相关方法
- 学习大数据量情况下的查询效率，以及不同数据建模方法的性能的对比
- 了解数据仓库、联机分析处理以及相关领域的研究方向

第一章　数据仓库基本概念

北京大学信息科学与技术学院

童云海

2005年3月

第一章 数据仓库基本概念

- 数据仓库技术产生的背景
- 什么是数据仓库技术
- 数据仓库技术与相关技术的比较和联系

信息技术发展的几个阶段

- 1960s: 数据采集、数据库创建阶段
  - 集中于原始文件的处理
  - 层次数据库和网状数据库
- 1970s: 关系数据库管理系统
  - 关系数据模型和关系数据库管理系统
  - E-R模型、SQL语言、查询处理和优化、OLTP（恢复和并发技术）
- 1980s: 高级数据库管理系统
  - 面向对象数据库、对象－关系数据库、主动数据库、演绎数据库、模糊数据库、空间数据库、时空数据库、统计数据库
  - 数据挖掘技术
- 1990s: 数据仓库、联机分析处理和数据挖掘
  - 数据仓库、联机分析处理和数据挖掘, 多媒体数据库, Web数据库、Data Stream

现有的数据库系统的侧重点

- 现有的数据库系统，主要用于事务处理
  - 一笔存款（一张存款单）
  - 一笔取款（一张取款单）
  - 一笔转帐（一张转帐单）
  - 一次挂失（一张挂失单）
- 强调多用户并发环境，数据的一致性、完整性

数据查询举例

- 查询2002年3月19日在工行北京分行海淀支行办理牡丹灵通卡挂失业务的客户资料
  - 数据库方法
    - （机构（机构名称=“工行北京分行海淀支行”）[机构代码]*卡资料表（卡状态=“挂失”^业务发生时间=“03/19/2002”^类别=“牡丹灵通卡”）[机构代码、客户号]*客户信息）[姓名，性别，单位，电话…]
  - 文件方法
    - 由应用程序实现，一段不小的程序（过程），包括打开、关闭文件，读、写一个记录

企业信息化建设现状

- 在数据库技术的支持下，一大批成熟的业务信息系统投入运行，为企业发展作出了巨大贡献
- 各类信息系统大多属于面向事务处理的OLTP系统
- 信息系统多年运行，积累了大量的数据
- 数据是一种宝贵的资源，但没有充分发挥作用
- 管理决策层对数据分析基础平台的需求日益强烈

信息化建设的趋势

- 发展趋势
  - 数据集中化
  - 业务综合化
  - 管理“扁平化”
  - 决策科学化
- 特点
  - 以客户为中心
  - 以服务求发展

企业信息化建设提出了更高的要求

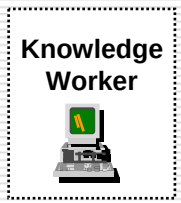
- 市场竞争日益激烈 — 创造竞争优势
  - 需要及时、准确的做出科学决策
  - 科学决策必须以准确、有效的数据为基础
  - 充分利用现有数据，将它转化为信息
- 以客户为中心的经营管理模式 — 优化客户关系
  - 原有系统往往以产品为中心
  - 原有系统往往以“单据（票证）”的处理为基础
  - 转向“以客户为中心”
  - 强调服务，尤其是个性化服务

分析处理的需求

- 例1：今年销售量下降的因素（时间、地区、商品、销售部门）
  - 时间：销售
  - 地区：（销售*顾客）[顾客地址所在的地区,……]
  - 商品：（销售*订单细则）[商品类别,……]
  - 销售部门：销售*员工*部门[部门名称,……]
- 例2：持卡人今年的交易情况与以往相比，有怎样的变化？交易特点（存款、取款、转帐、消费）是什么？持卡人消费倾向（宾馆、大型商场、超级市场等）是什么？
- 要求：
  - 多个子系统中的数据（数据集成）
  - 历史数据
  - 汇总、综合的数据
  - 一致的数据视图

分析人员典型的信息需求

- 覆盖企业内部信息、合作伙伴信息和市场信息
- 覆盖综合信息和明细信息
- 覆盖当前数据和历史数据
- 高可用性
- 高质量的数据（一致性、完整性）
- 支持各种不同的分析方法
- 数据定义符合业务人员要求



分析决策人员的挑战

组织内部

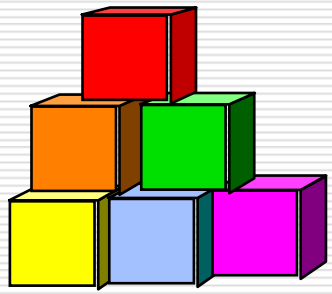
- 横向共享信息
- 数据的重构
- 个人授权
- 服务和质量管理

组织之间

- 合作伙伴
- 客户驱动的解决方案
- 战略联盟
- 价值链和供应链

市场

- 竞争对手
- 市场分割
- 实时的市场行情
- 全球化



现有数据库系统处理分析型应用存在的问题

- 数据可信性
- 生产率
- 不可能把数据转换成信息
- 数据动态集成问题
- 历史数据问题
- 数据的综合问题：非细节数据，多种程度的综合

数据可信性

- 数据没有同一时间基准
  - 例如：一个企业的两个部门向管理者呈送报表
    - 部门A，于星期天傍晚抽取了分析所需的数据，结论为业绩上升10%
    - 部门B，于星期三下午抽取了分析所需的数据，结论为业绩下降15%
- 算法不同

  部门A使用的是旧帐号

  部门B使用的是大帐号
- 多次抽取，扩大了上述两个问题

  用抽取程序从数据库/文件中抽取数据，并存放起来，然后又在此基础上再次进行抽取，从数据进入系统到提供分析往往经过8、9次的抽取。

数据可信性（续）

- 外部数据问题
  - 一位分析员把《华尔街日报》的数据带进系统
  - 另一位将《商业周刊》的数据进入系统
  - 数据一旦进入系统，往往已失去“身份”，并且一位分析员也不知道另一位分析员所输入的数据
- 开始时就不是同一个公共的数据源
  - 部门A最初来源于文件XYZ
  - 部门B最初来源于DB　ABC

生产率

- 为了生成一个企业报表，必须经过
  - 获得源数据
  - 定位和分析数据：由于同名不同义、同义不同名，很难准确定位和分析，可能造成进一步的混乱
  - 把数据加工成报告
    - 要写许多程序，每个程序必须客户化（与客户环境有关）
    - 程序会涉及公司具有的各种技术
    - 由于定位数据困难，检索所要的数据是一件很麻烦的事
  - 完成任务需要很长时间
    - 定位数据 + 获得数据 + 集成报告，完成任务所需时间较长
    - 每份报告各自需求不同，因此每份报告所需要的时间都很长。

从数据到信息

- 例如：“今年的帐户情况与前五年比较”
  - 涉及大量应用：储蓄应用、贷款、即期汇票管理、信托，而这些应用并未集成。
  - 没有足够的历史数据：
    - 贷款部门，拥有二年的数据
    - 银行存折处理，拥有一年的数据
    - 即期汇票管理只有60天的数据
    - 现金交易处理具有18个月的数据。
  - 数据不一致问题：同名不同义、同义不同名，例如M/F，Male/Female
  - 外部数据和非结构化数据

操作型环境和分析型环境

- 不同的需求，要求将操作型环境和分析型环境相分离
  - 在操作型环境中支持分析应用太复杂、太困难
  - 操作性环境不支持域（Domain）之间的联系，仅仅支持表之间的连接
  - 不同的数据环境要求从数据组织（结构）和操作上进行工作

两种数据的区别

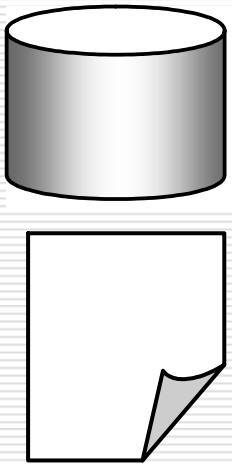
原始数据/操作型数据	导出数据/分析型数据
面向应用	面向主题
详细的	综合的，或提炼的
在访问瞬间是准确的	代表过去的数据，快照
是为日常工作服务	为管理者服务
可更新	不更新
重复运行	启发式运行
处理需求预先可知	处理需求事先不知道
非冗余性	总是存在冗余
对性能要求高	对性能要求宽松
一次访问一个单元	一次访问一个集合
静态结构：可变的内容	结构灵活
访问频繁	访问很少或不多

体系结构化环境的层次

操作层	原子/数据仓库层	部门层	个体层
1、细节的	1、大部分是粒度化数据	1、领域狭隘	1、暂时的
2、日常的	2、反映历史变化	2、一些导出数据和一些原始数据	2、为特定目的的
3、当前值的	3、集成的	3、典型的部门 如：财务 市场 工程 保险 制造	3、启发式的
4、访问频繁	4、面向主题		4、非重复的
5、面向应用	5、一些汇总		5、基于PC和工作站的

两种报表的区别

操作型

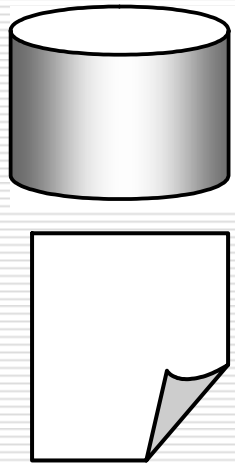


操作型报表

1、主要是行式项目；即使有综合的，也很少或不重要

2、对办事层人员是重要的

数据仓库



数据仓库报表

1、即使有行式项目也很少甚至没有用；综合或其它计算非常重要

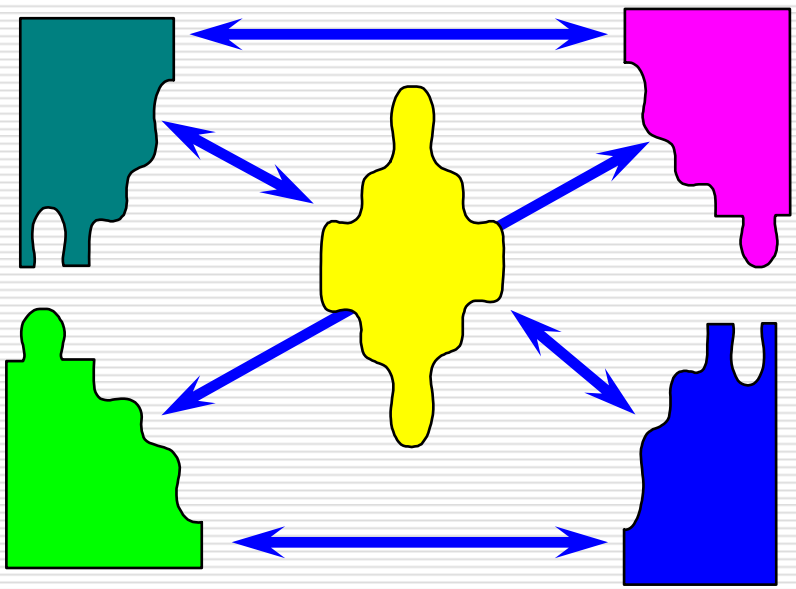
2、对管理层人员是重要的

两种报表的区别（续）

例：就一个银行而言

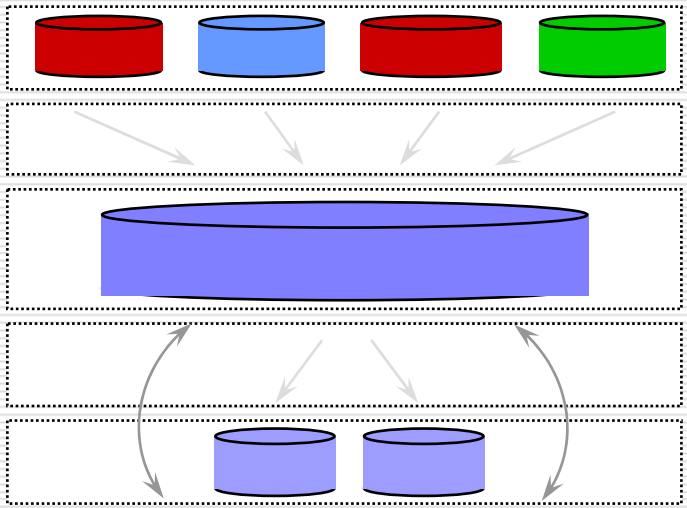
- 出纳员需要操作型报表，因为他需要知道当天所有交易，来确定一天结束时的现金余额；
- 银行行长的长期战略决策（如决定一个地区安装ATM机的数目）就需要了解大量的内部和外部信息，每天的交易报表对他意义不大，他更需要分析型报表

提升现有的信息



- 企业范围内的信息共享
- 准确、一致的集成数据
- 面向整个企业和最终用户，针对分析需要，进行数据重组，形成一套全新的、相对完整的数据视图
  - 快速访问
  - 精确、灵活分析

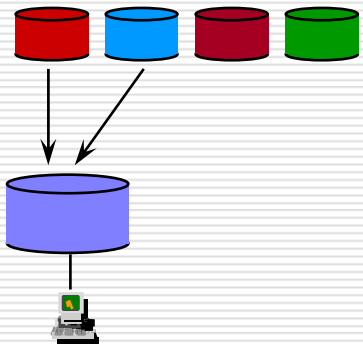
数据仓库要解决的基本问题

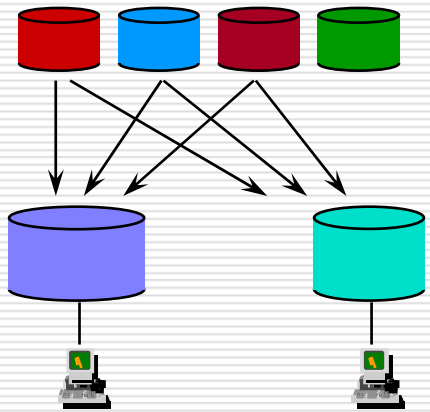


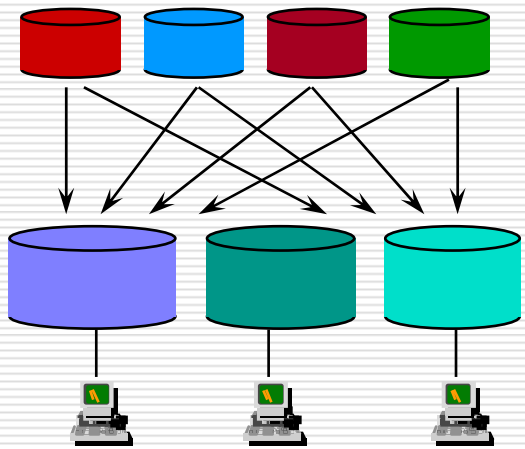
YOUR
D A T A

- 全局范围内统一数据视图
  - 数据内容
    - 数据的完整性
    - 数据的准确性
    - 数据的一致性
  - 数据组织
    - 面向分析决策

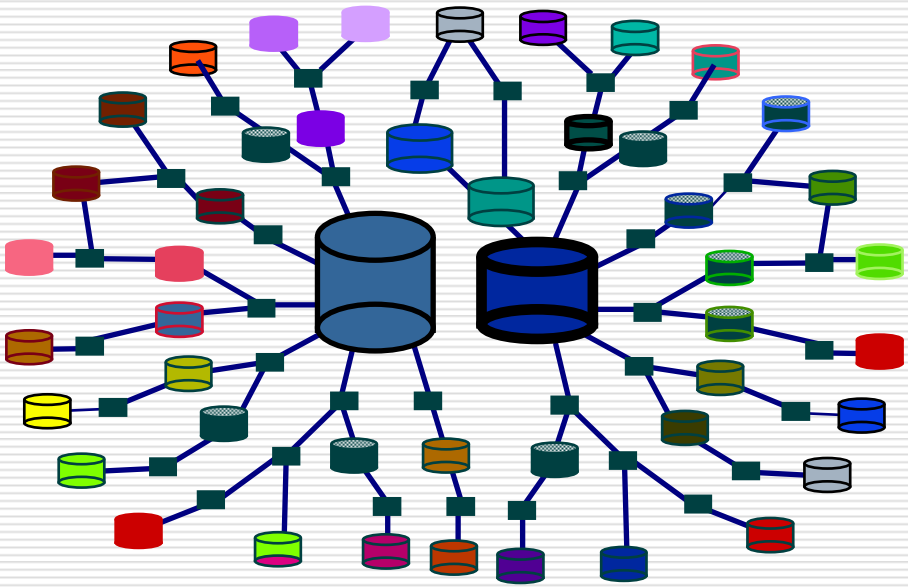
在实际中经常存在这样...







“蜘蛛网”问题



- 没有统一规划和设计
- 数据模型不一致
- 数据定义不一致
- 数据准确性差，冗余度高
- 业务流程发生变化
- 历史数据不统一、不规范

解决方案：

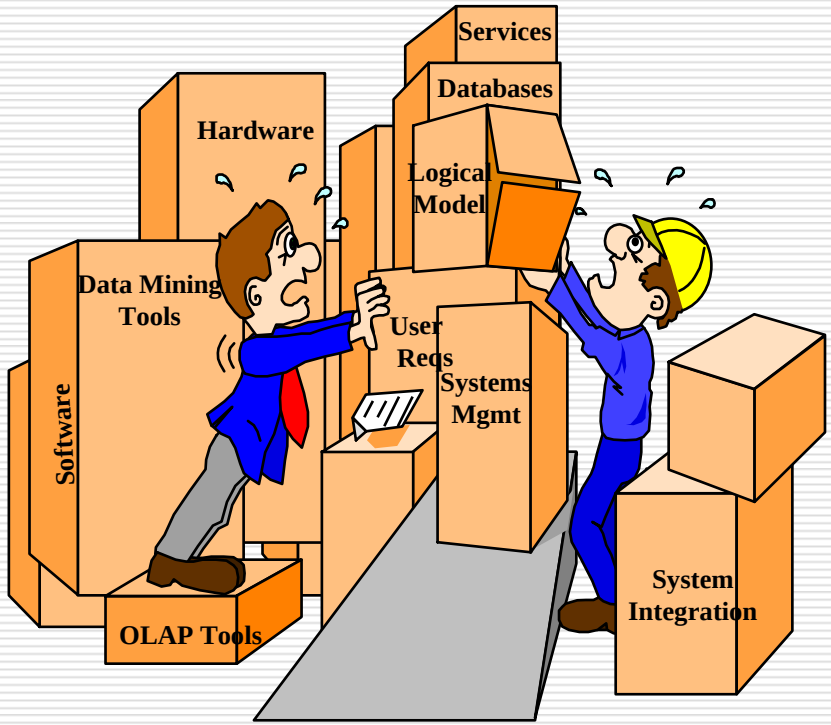
深入、全面、客观的数据源分析

建立数据仓库系统

Source: "Building the Data Warehouse," W. Inmon, QED Publishers

数据仓库需要建立，而不是购买

- 需要针对多个数据源的数据集成
- 考虑“重要”的业务分析问题
- 选择合适的数据源（内部、外部）
- 数据仓库系统的建设永无止境
- 数据仓库系统的建设是一项工程，同时也是一个过程



第一章 数据仓库基本概念

- 数据仓库技术产生的背景
- 什么是数据仓库技术
- 数据仓库技术与相关技术的比较和联系

数据仓库的定义

A single integrated store of data which provides the infrastructural basis for informational software applications ***in the enterprise***

The place to publish corporate or organization data which:

- **is consistent and accessible**
- **allows separation or combinations to measure business**
- **has a set of query, analysis and presentation tools**

The process of turning raw data into information so users can…

- **make tactical and strategic decisions**
- **make better decisions faster**
- **capitalize on opportunities**

数据仓库的定义

- 数据仓库（Data Warehouse）是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant） 的数据集合，用于支持管理决策和信息的全局共享。

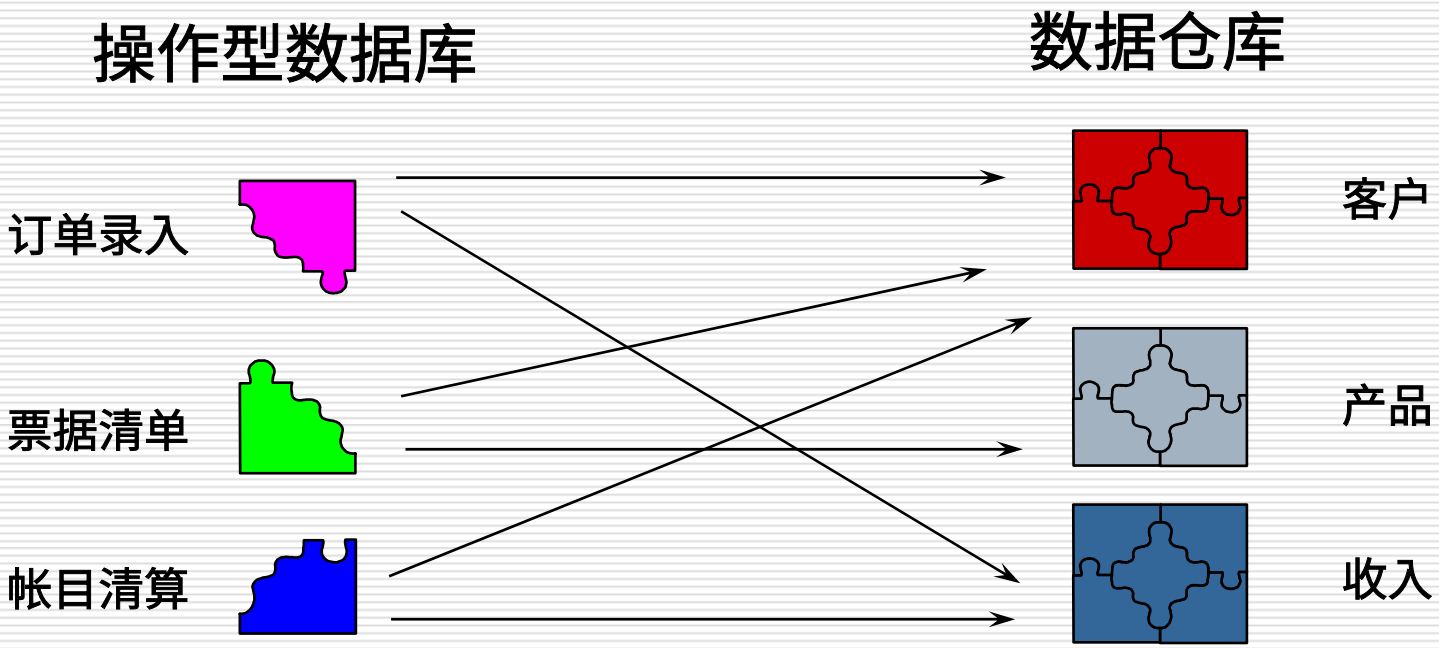
  —— W.H.Inmon

- 对数据仓库的理解
  - 数据仓库用于支持管理和决策，面向分析型数据处理，它不同于企业现有的面向交易的操作型数据库；
  - 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。
- Data warehousing:
  - 构建和使用数据仓库的过程

数据仓库回答的问题

- 数据仓库技术将为高层管理人员的科学决策提供可靠依据。
  - 去年各个地区各个产品的销售量和销售额？
  - 10年以来，各个计算机厂商每个季度的销售额占有比例的变化情况？
  - 如果某种产品的销售价格打9折，利润将发生怎样的变化？
  - 今年销售量下降的主要因素（时间、地区、商品、销售部门）是什么？

数据仓库的特点：面向主题



- 操作型数据库是面向特殊处理任务，进行组织，由各个不同的系统独立维护
- 数据仓库是面向不同的主题域进行组织。一个主题通常与多个操作型信息系统相关

面向主题的数据组织

- 主题：宏观分析领域所涉及的分析对象
  - 面向主题的数据组织方式：在较高的层次上对分析对象的数据的一个完整、一致的描述。
  - 采用面向事务进行数据组织（见前面例子），其特点为：
    - 充分考虑企业的部门组织结构和业务活动
    - 反映企业内部数据流动情况，业务处理的数据流程
    - 与业务处理流程中的单据、票证、文档有良好的对应
    - 数据与应用（数据的处理）有一定的对应
  - 例：保险公司：

    面向应用（操作）：财产险、寿险、健康险、意外险。

    面向主题：客户、保单、保费、理赔（赔款）。

面向主题的数据组织的特点

- 各个主题有完整、一致的信息内容，便于在此基础上作分析处理。
- 主题之间有重迭的内容，反映主题间的联系。
- 重迭是逻辑上的，不是物理上的；重迭仅在细节层。
- 各主题的综合方式不同。例如：
  - 商品主题的采购信息可汇总（综合）成：商品号、时间段、采购总量… …
  - 供应商主题的供应商品信息可综合成供应商号、时间段、供应总量… …
- 主题域应该具有独立性、完备性。
  - 独立性：有明确界限，数据是否属于该主题；
  - 完备性：对该主题进行分析所涉及的内容均要在主题域内；

面向主题数据组织的实现

- 多个表，公共码键（把各个表统一联系起来），但同一主题的表可存放在不同介质上

例：商品主题可有商品表（商品基本信息），采购表（商品采购信息），销售表（商品销售信息），库存表（商品库存信息）； 公共码键：商品号。

- 综合信息，多个层次
- 面向主题数据组织方式独立于数据的事务处理逻辑。即可以支持分析型数据环境，又可用于ODS（操作数据存储）系统（作为全局数据库的数据组织方式）

面向主题数据组织的实现（续一）

- 多个表

基本顾客数据
1985－1987

顾客ID
起始日期
终止日期
姓名
地址
电话
出生日期
性别
……

基本顾客数据
1988－1990

顾客ID
起始日期
终止日期
姓名
地址
信用度
雇主
出生日期
性别
……

顾客活动
1986－1989

顾客ID
月份
  交易数目
  平均交易额
  最高交易额
  最低交易额
  已取消交易数
  ……

顾客活动细节
1987－1989

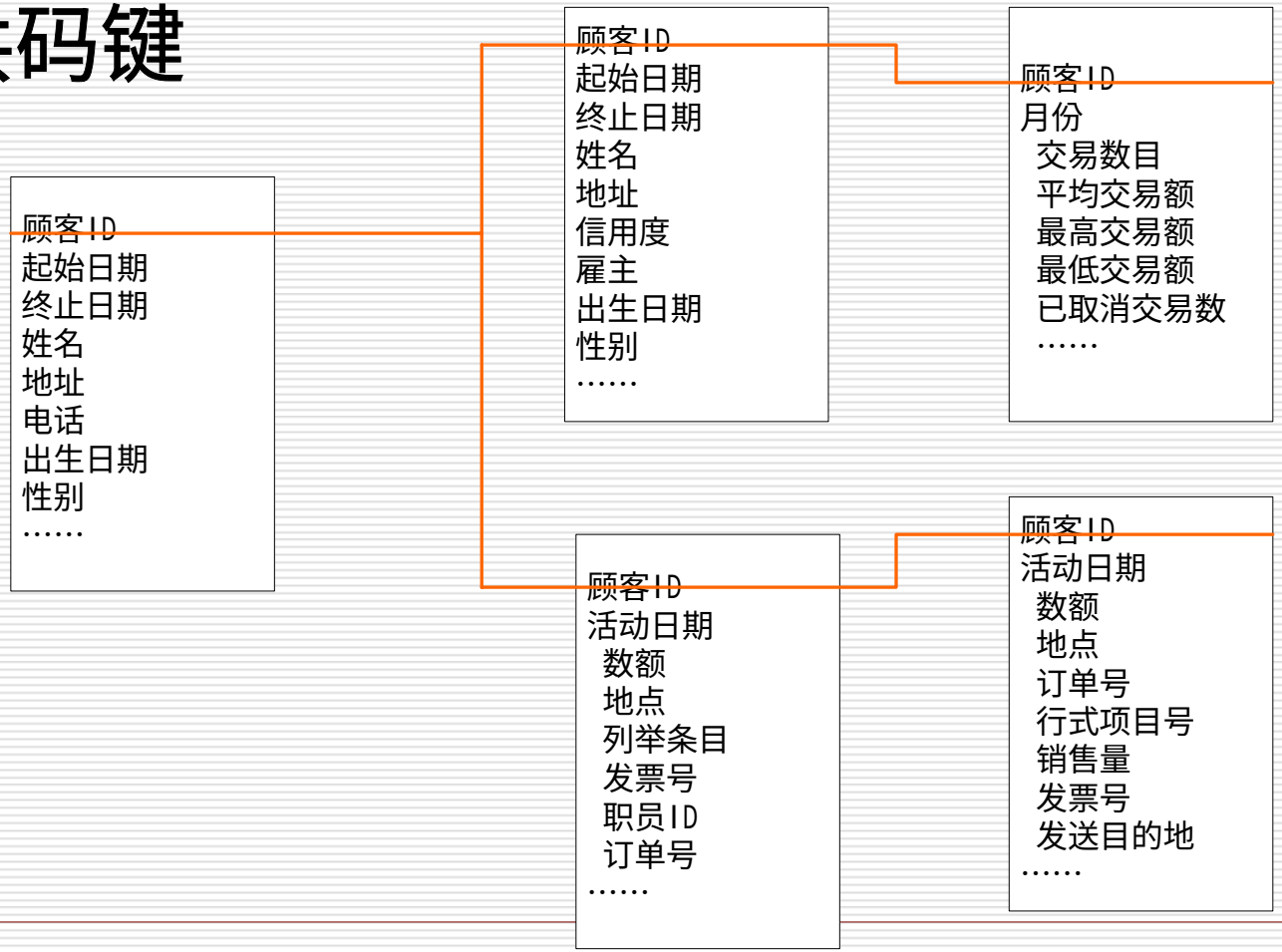
顾客ID
活动日期
  数额
  地点
  列举条目
  发票号
  职员ID
  订单号
……

顾客活动细节
1990－1991

顾客ID
活动日期
  数额
  地点
  订单号
  行式项目号
  销售量
  发票号
  发送目的地
……

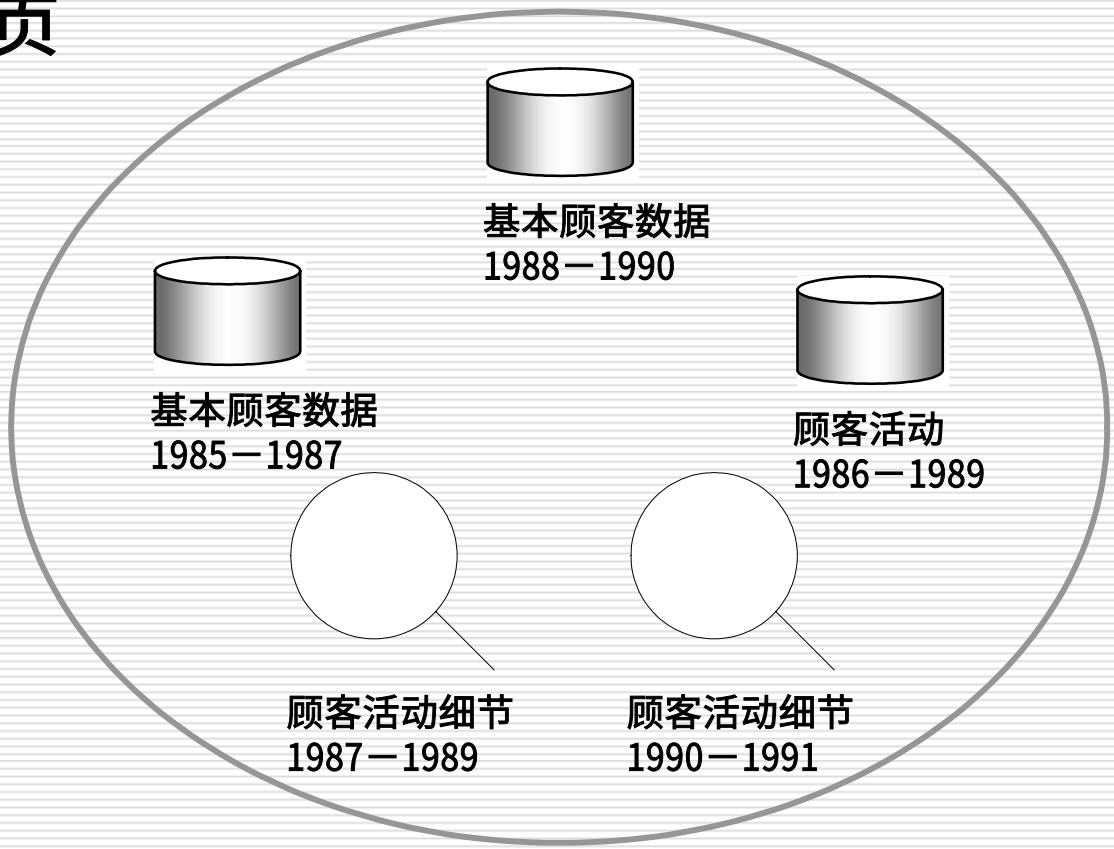
面向主题数据组织的实现（续二）

- 公共码键



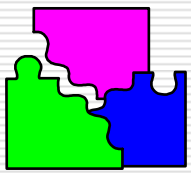
面向主题数据组织的实现（续三）

- 不同介质



数据仓库的特点：集成的

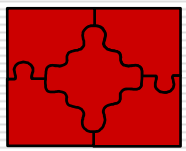
操作型数据库



面向特殊应用

- 每一个数据库面向特定的应用，各类应用（包括其相关的数据库）之间相互独立。
- 系统的发展经历一个长期的过程

数据仓库



集成的

- 数据仓库中的数据从建立时开始，面向整个企业的分析处理，数据仓库中的数据是已经集成了的，消除了数据的不一致性
- 在某个时间点完成设计，实现需要经历一个长期的不断迭代的过程

数据仓库的特点：集成的

- 消除冲突：不一致，同名异义、异名同义、单位不统一等等，需要进行数据清理(因为来源于不同的子系统，与不同的主要逻辑捆绑)

编码 $\left.\begin{array}{l} m, f \\ 1,0 \\ x, y \\ male, female \end{array}\right\} \rightarrow m, f$

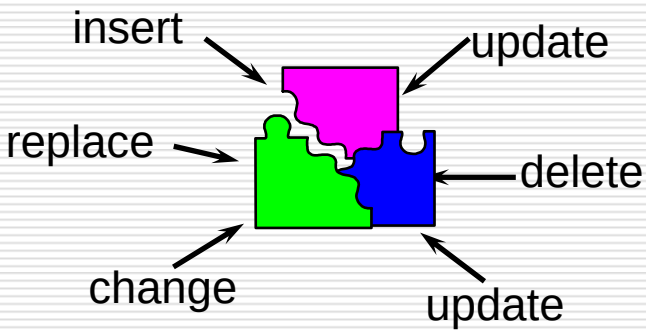
属性度量 $\left.\begin{array}{l} cm \\ inches \\ yds \end{array}\right\} \rightarrow cm$

码描述 $\left.\begin{array}{l} char(10) \\ decfixed(9,2) \\ pic'9999999' \\ char(12) \end{array}\right\} \rightarrow Char(12)$

- 数据的综合和计算:可在抽取数据时；也可在进入DW以后。

数据仓库的特点：相对稳定的

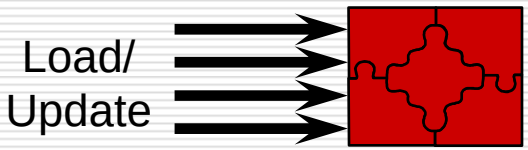
操作型数据库



实时更新

- 随时更新
- 数据根据需要进行变化，并不是按照一定周期进行修改

数据仓库



在某个时间点保持不变

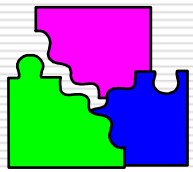
- 定期加载，加载后的数据极少更新。
- 并不意味着数据仓库中的数据不更新

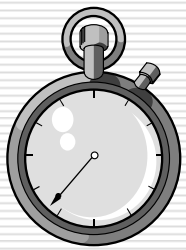
数据仓库的特点：相对稳定的

- 一般不修改，只追加；过期限的数据可从DW中移走（删去）；
- 对DW，主要是查询，DWMS比DBMS要简单
  - 可不考虑并发控制
  - 要考虑性能（因为查询数据量大）和界面友好（对高层管理者）

数据仓库的特点：反映历史变化

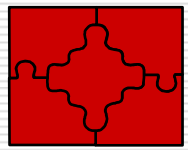
操作型数据库

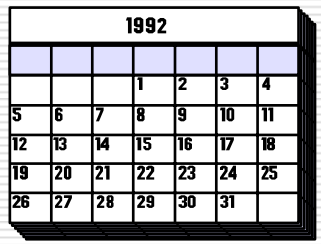




- 主要关心当前数据

数据仓库





- 通常关心历史数据

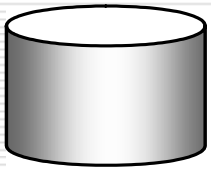
数据仓库的特点：反映历史变化

- 码键包含时间项
- 不断增加新的数据内容；
- 删去过时的数据；例如：超过10年的数据
- 与时间有关的综合数据：随时间变化而重新组合

数据仓库的特点：反映历史变化

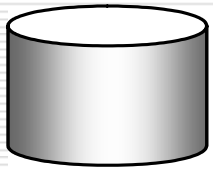
- 操作型数据与DW中的数据比较

操作型环境



- 60-90天数据
- 记录能被更新
- 码中不一定包括时间元素

数据仓库



- 5-10年数据
- 数据的复杂快照
- 码中包括时间元素

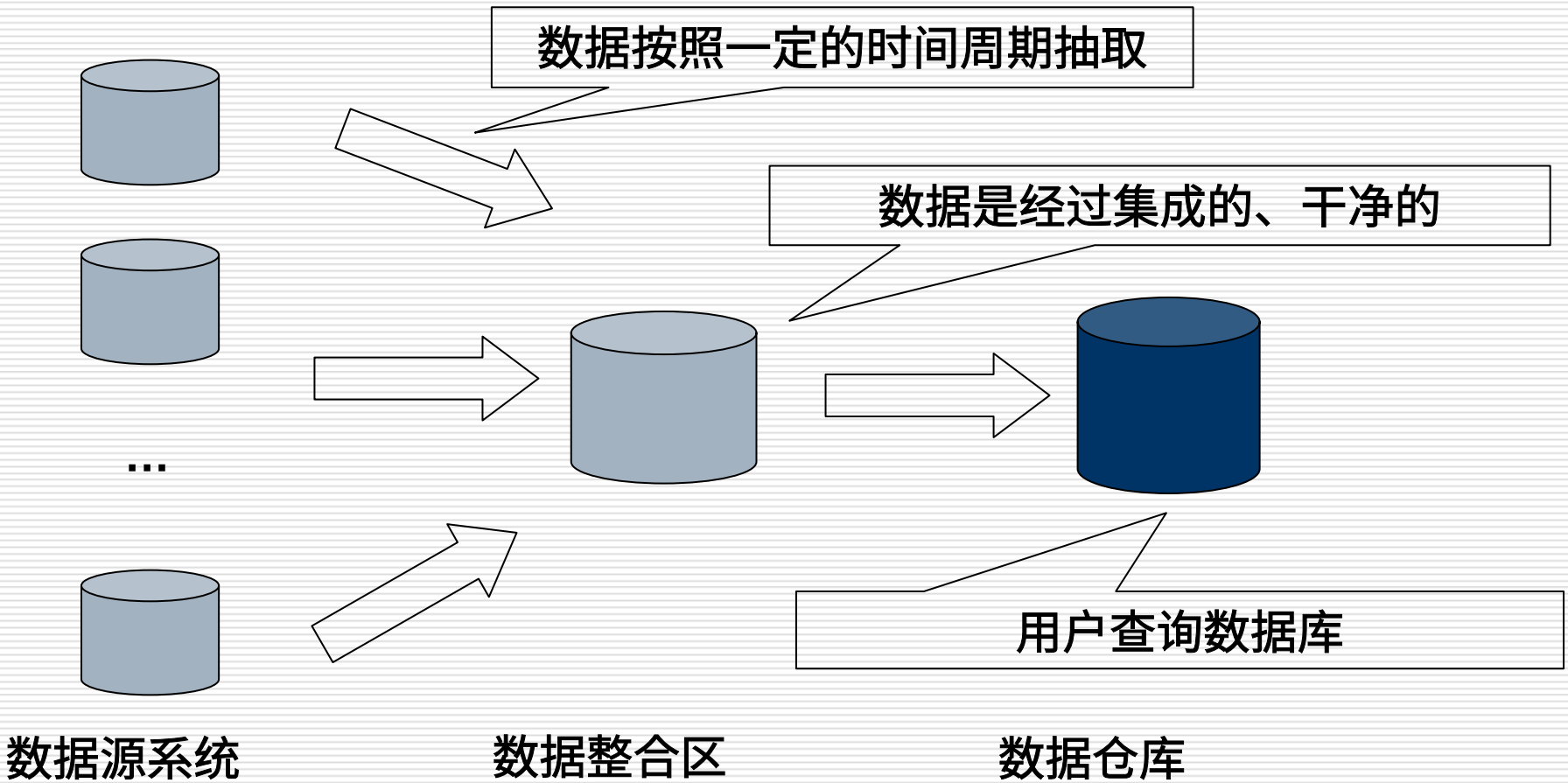
第一章 数据仓库基本概念

- 数据仓库技术产生的背景
- 什么是数据仓库技术
- 数据仓库技术与相关技术的比较和联系

数据库技术与数据仓库技术

- 数据库技术在系统功能和性能需求
  - 强调的是多用户环境下如何针对并发用户的增删改操作，保证数据的一致性和可恢复性，并发用户的吞吐量为数据库管理系统的重要性能指标
- 数据仓库技术在系统功能和性能需求
  - 强调的是大数据量环境下的高效、快速查询，查询的吞吐量为数据仓库管理系统的重要性能指标

数据的抽取、转化和加载

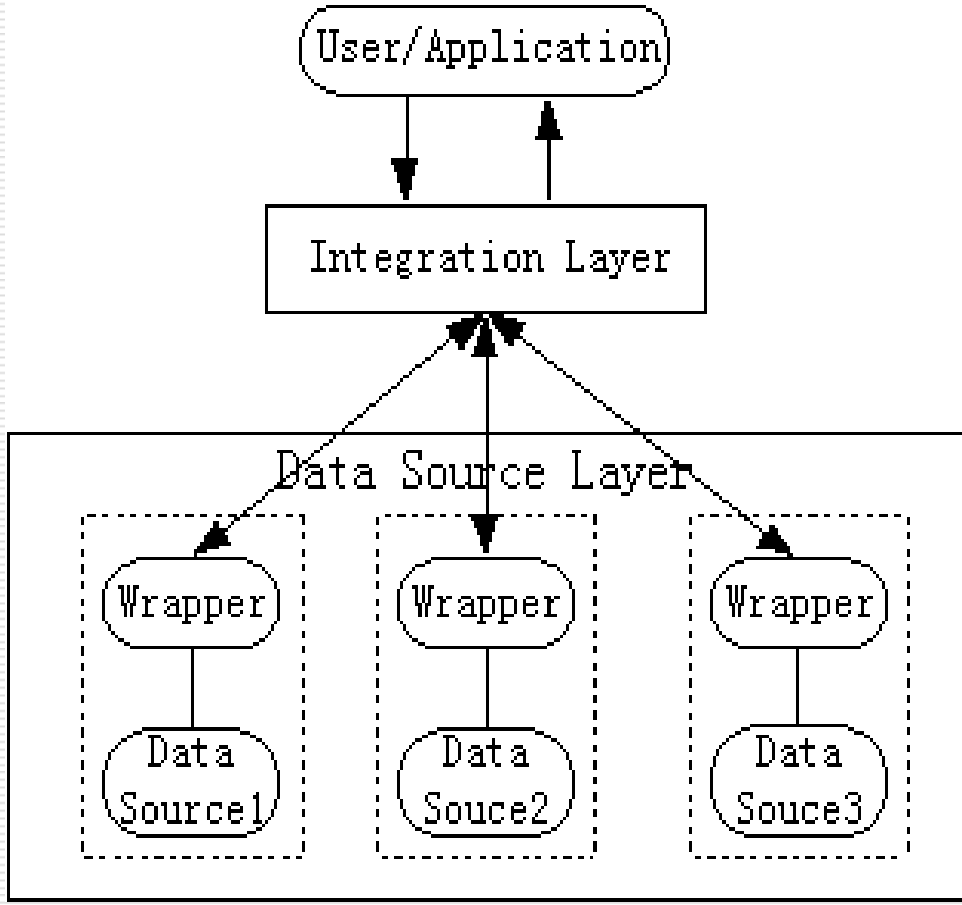


数据集成是大问题

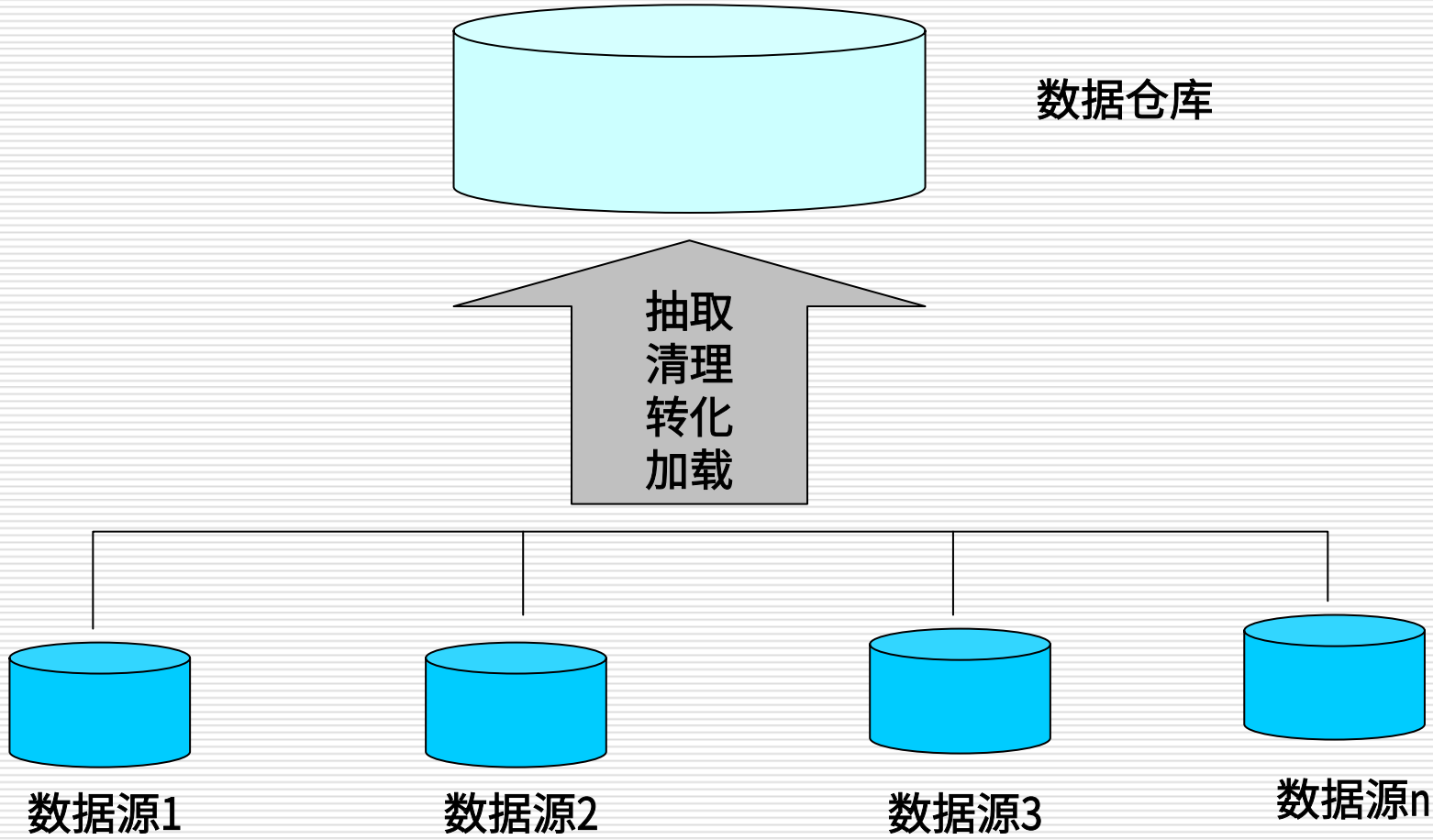
- 数据仓库是多个数据源数据的综合
- 数据必须转换成一个一致的格式
- 对于一个典型的数据仓库系统建设项目中，数据集成工作通常占到整个系统建设的80%
- 集成困难的原因:
  - 缺乏元数据或者根本就不存在
  - 数据质量很差
    - 存在大量的空缺值
    - 存在大量的同名异义或者同义异名的问题
  - 不一致的语义
  - ……

数据集成的方法: MQS

- MQS：Mediated Query System
  - 1992年，由Stanford University的Gio Wiederhold
  - 查询驱动的方法
  - 其目标是实现对信息的智能、能动的使用
  - Mediator是一个软件模块，实现对数据的抽象与表示，具有相当的智能
  - Mediator具有某些数据集的知识，为高层应用服务
  - Mediator本身还可以进一步抽象成MetaMediator，来描述关于Mediator的信息

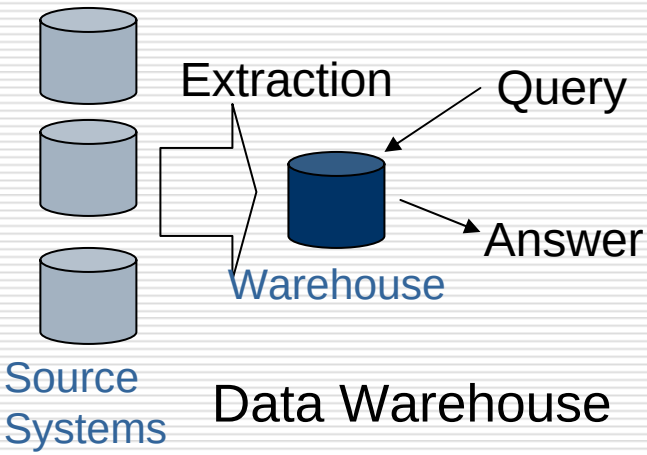


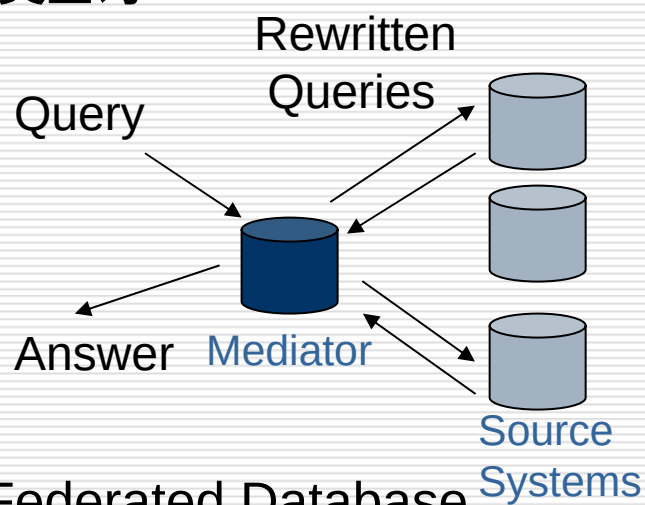
数据仓库的数据集成



联邦数据库

- 数据仓库
  - 对于所有数据创建一个备份
  - 基于备份上重构的数据，执行分析查询
- 联邦数据库
  - 从数据源中检索所需要的数据以回答各类查询
- “lazy” vs. “eager” 的数据集成





数据仓库 vs. 联邦数据库

- 联邦数据库的优点
  - 不需要冗余数据的拷贝
  - 查询的结果反映所涉及数据的实时情况
  - 安全策略更加方便
- 联邦数据库的缺点
  - 分析查询对于事务系统增加了额外的“Load”数据的开销
  - 查询优化很难做得很好
  - 历史数据可能不存在或者不可用
  - “wrappers”的功能很复杂，需要在分析服务器和数据源系统之间进行沟通
- 在实践中数据仓库方法变得更加普遍
  - 更好的性能
  - 更低的复杂性
  - 对于分析来说，缺少部分实时的数据是可以接受的