数据仓库技术

谭义红：15873193369；　　QQ：2647724
yhtan09@163

课程介绍

- 性质
  - 是信科专业的主要专业课程、是决策支持系统方向的重要课程
- 目的
  - 理解数据仓库及OLAP的相关概念
  - 了解数据仓库及OLAP的发展趋势和应用领域
  - 掌握数据仓库的设计、构建，数据的准备、转换、装载，数据的浏览、分析等方法和技术。
- 相关课程
  - 数据库、数据挖掘、决策支持系统设计与开发

为什么学本课程（Why）

- 信息技术在商业中发展
  - 管理信息系统（MIS）
  - 企业资源计划系统（ERP）、客户关系管理（CRM）
  - 商业智能系统（BI）
- 发展过程中存在的问题
  - 数据可信性
  - 生产率问题
  - 无法将数据转化为信息

为什么学本课程（Why）

- **BI**
  - 定义（IBM）：商业智能是一系列由系统和 技术支持的以简化信息收集和分析的策略集合，它应该包括企业需要收集什么信息、谁需要去访问这些数据、如何把原始数据转化为最终战略性决策的智能、客户服务和供应链管理。
  - 包括：数据仓库（DW）、联机分析（OLAP）、数据挖掘（DM）
  - 工具：IBM、Oracle、Microsoft、SAS、CA等

为什么学本课程（Why）

- 市场需求（岗位）
  - 数据仓库工程师
    - 岗位要求1、岗位要求2、岗位要求3
  - 数据仓库开发工程师
    - 岗位要求1
  - 数据仓库BI架构师
    - 岗位要求
  - 数据仓库高级开发工程师
    - 岗位要求
  - 数据仓库测试工程师
    - 岗位要求
  - 数据仓库咨询师
    - 岗位要求

本课程的主要内容（What）

- 数据仓库与**OLAP**的相关理论知识
- 数据仓库设计
- 数据准备、转换、装载（**SSIS**）
- 多维数据集操作及分析**(SSAS)**
- 多维数据分析报表**(SSRS)**

如何学好本课程（How）

- 重视相关概念和原理的理解
- 从全局把握上把握数据仓库创建、管理及OLAP分析技术框架
- 从微观角度掌握具体技术细节
- 主动、认真做好实验及课程设计

教材：

《数据仓库设计：现代原理与方法》

（美）Matteo Golfareli著

参考教材：

1） (美)JOY MUNDY.数据仓库工具箱--面向SQL SERVER 2019和MICROSOFT商业智能工具集. 北京：清华大学出版社,2019

2)于宗民,刘义宁, 祁国辉.数据仓库项目管理实践.北京：人民邮电出版社2019

• 朱德利. SQL Server 2019数据挖掘与商业智能完全解决方案.北京：电子工业出版社，2019.

• technet.microsoft/zh-cn/default.aspx

3) msdn.microsoft/zh-cn/sqlserver/default.aspx

4) dwway/html/news.html

第1章 数据仓库与OLAP概述

1.1 决策支持系统

- 决策支持系统
  - **DSS(decision support system)是可扩展交互式IT技术和工具的集合，这些技术和工具用于处理和分析数据以及辅助管理人员制定决策。为此，这种系统匹配管理人员的个人资源和计算机资源，以提高决策质量。**

1.2 数据仓库

❖数据仓库的引入

了解各销售员的月销售情况

商品名称	生产厂家	销售时间	销售地	销售员	销售量
空调	美的	2009.3.12	北京	001	2
空调	格力	2009.5.3	长沙	002	3
空调	美的	2009.5.10	北京	001	2
空调	格力	2009.6.3	长沙	002	3
空调	美的	2009.7.10	北京	001	2
空调	格力	2009.8.3	长沙	002	3
空调	美的	2009.10.10	北京	001	2
空调	格力	2009.11.3	长沙	002	3
电视机	TCL	2009.10.3	长沙	003	3

了解格力空调在长沙的月销售情况 了解各品牌空调在各城市的月销售情况

1.2 数据仓库

❖数据仓库的引入

客户代码	支行管辖机构...	贷款类别明...	贷款期限明...	借款日期	贷款总额
80922925...	13080310101	I158	24	2000-12-22 0:0...	2305821.00
170067234...	13063011006	A101	11	2002-3-27 0:00:00	6874392.00
170077081...	13091511309	A101	11	2001-4-9 0:00:00	2510000.00
170077081...	13091511309	A101	11	2001-4-9 0:00:00	1720000.00
110777133...	13010110104	A101	11	2000-4-24 0:00:00	358000.00
170081667...	13070120901	A101	21	2002-7-1 0:00:00	7555821.00
110896110...	13070210105	A101	11	2000-11-13 0:0...	1800000.00
113e00150...	13080112103	C113	11	2000-4-19 0:00:00	100000.00
172160678...	13300710203	A101	21	2002-4-28 0:00:00	7024392.00
172160678...	13300710203	A101	21	2002-4-28 0:00:00	6024392.00

1. 2 数据仓库

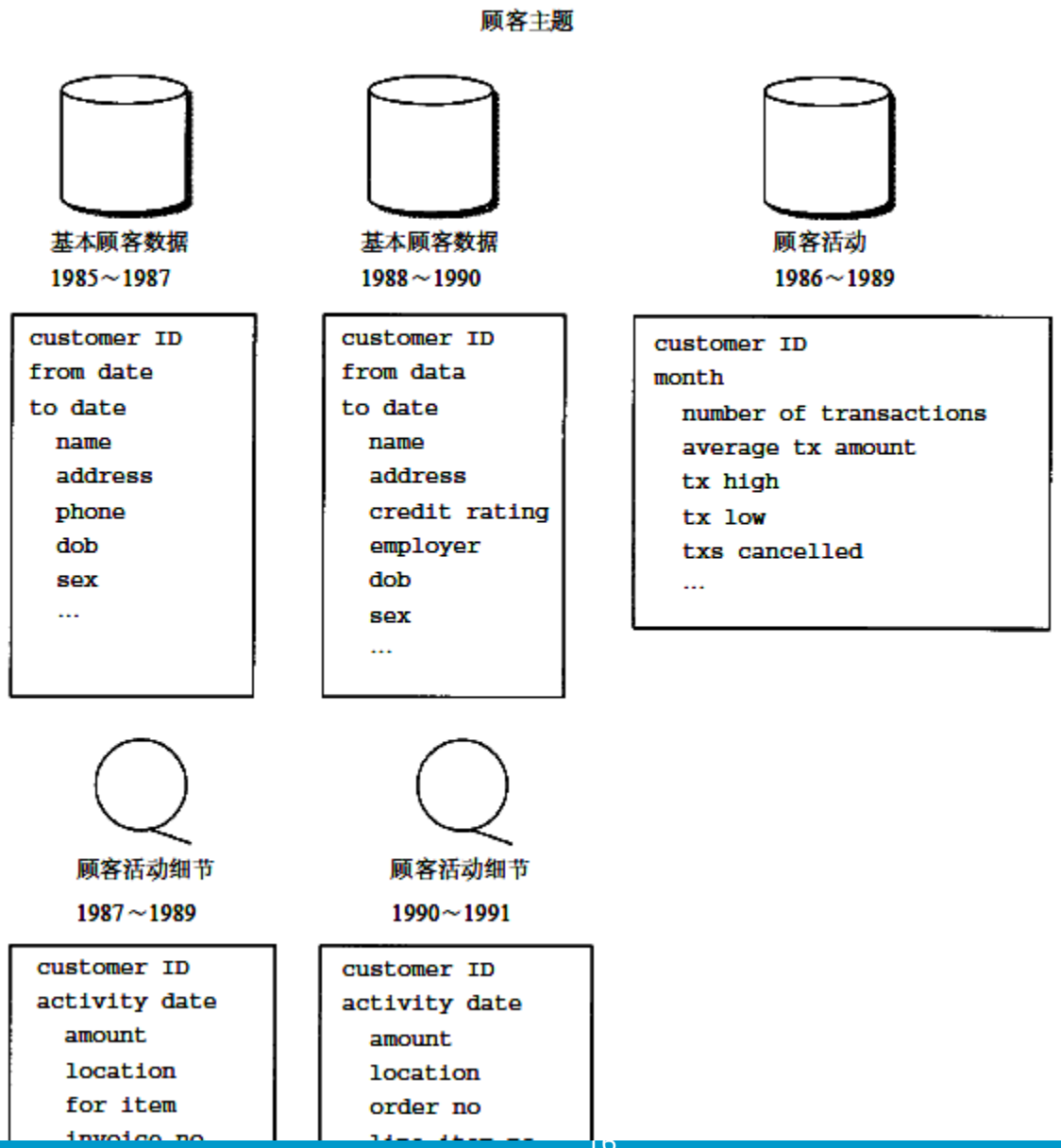
❖数据仓库的引入

	市行名称 ▾	贷款期限名称 ▾					
	⊟A市				⊞B市	⊞C市	⊞D市
	长期贷款	短期贷款	中期贷款	汇总			
贷款类别名称 ▾	贷款总额	贷款总额	贷款总额	贷款总额	贷款总额	贷款总额	贷款总额
房地产贷款	32305821	24316034	146252389	202874244	217830278	469415971.48	516020775
扶贫贷款		516985844	52635673	569621517	104577241		1750000
固定资产贷款	341140747	6805821	59364031	407310599	375602763.51	1641248405	790039306.95
流动资金贷款		12948854395.28	1759043473	14707897868.28	9706605201.46	32165204464.12	15935466279.6
贸易融资贷款		263041263		263041263	550271469.88	2223852654.84	508549677
票据贴现		5417079269.76		5417079269.76	17674301187.29	99381270094.19	14691408962.0
其他贷款类		8566803		8566803	20511642	19723284	152310531
汽车贷款							
生产经营类贷款							
一般消费贷款		10000		10000			5019192
总计	373446568	19185659430.04	2017295566	21576401564.04	28649699783.14	135900714873.63	32600564723.5

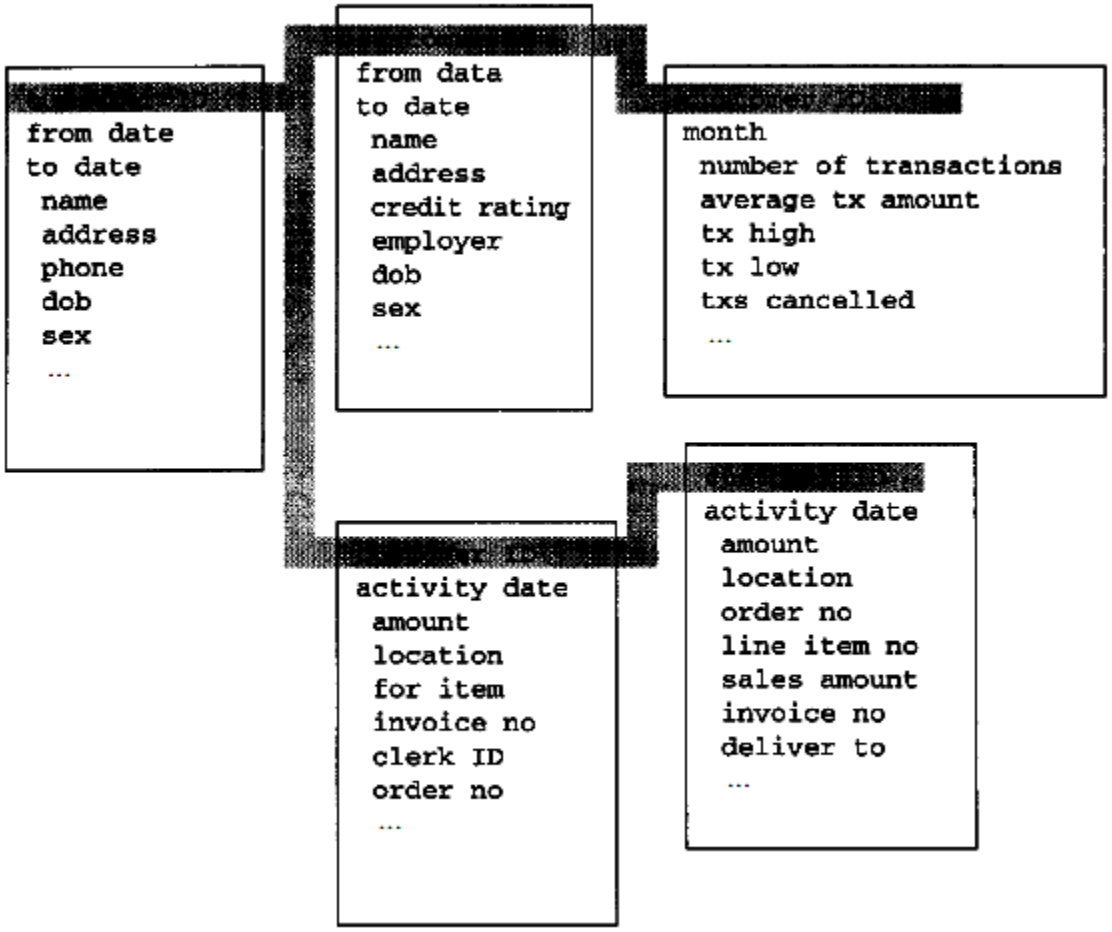
1.2 数据仓库

- 数据仓库的概念和特点
  - 概念：
    - 数据仓库（DataWarehouse，DW）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，它用于支持企业或组织的决策分析处理。
  - 特点：
    - 面向主题
      - 主题是在较高层次上将企业信息系统中的数据综合、归类和分析利用的抽象概念。每个主题对应一个分析领域。
      - 典型主题：客户、产品、销售、利润、保险等
      - 每个主题域都是以一组相关的表来具体实现，通过公共关键字建立联系

1.2 数据仓库


顾客主题
基本顾客数据
1985～1987
customer ID
from date
to date
name
address
phone
dob
sex
...
基本顾客数据
1988～1990
customer ID
from data
to date
name
address
credit rating
employer
dob
sex
...
顾客活动
1986～1989
customer ID
month
number of transactions
average tx amount
tx high
tx low
txs cancelled
...
顾客活动细节
1987～1989
customer ID
activity date
amount
location
for item
顾客活动细节
1990～1991
customer ID
activity date
amount
location
order no

1.2 数据仓库


from date
to date
name
address
phone
dob
sex
...
from data
to date
name
address
credit rating
employer
dob
sex
...
month
number of transactions
average tx amount
tx high
tx low
txs cancelled
...
activity date
amount
location
for item
invoice no
clerk ID
order no
...
activity date
amount
location
order no
line item no
sales amount
invoice no
deliver to
...

1.2 数据仓库

●各子系统建立数据库情况

子系统	数据库名称	数据字段
销售子系统	顾客	顾客号，姓名，性别，年龄，文化程度，地址，电话
	销售	员工号，顾客号，商品号，数量，单价，日期
采购子系统	订单	订单号，供应商号，总金额，日期
	订单细则	订单号，商品号，类别，单价，数量
	供应商	供应商号，供应商名，地址，电话
库存管理子系统	领料单	领料单号，领料人，商品号，数量，日期
	进料单	进料单号，订单号，进料人，收料人，日期
	库存	商品号，库房号，库存量，日期
	库房	库房号，仓库管理员，地点，库存商品描述
人事管理子系统	员工	员工号，姓名，性别，年龄，文化程度，部门号
	部门	部门号，部门名称，部门主管，电话

1.2 数据仓库

- 面向主题的数据组织

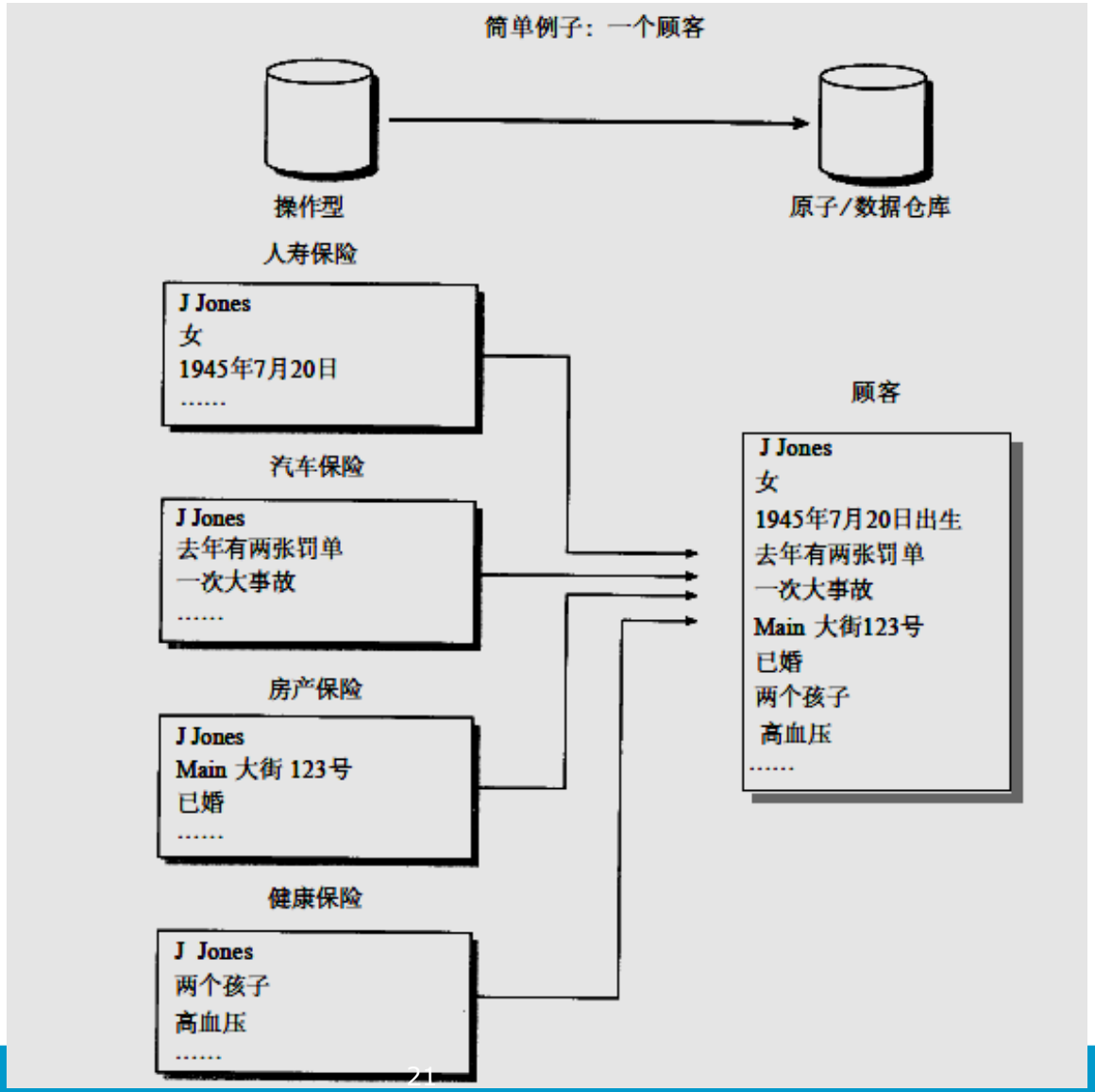
主题	信息类	数据字段
商品	商品固有信息	商品号, 商品名, 类别, 颜色
	商品采购信息	商品号, 供应商号, 供应价, 供应日期, 供应量
	商品销售信息	商品号, 顾客号, 售价, 销售日期, 销售量
	商品库存信息	商品号, 库房号, 库存量, 日期
供应商	供应商固有信息	供应商号, 供应商名, 地址, 电话
	供应商品信息	供应商号, 商品号, 供应价, 供应日期, 供应量
顾客	顾客固有信息	顾客号, 顾客名, 性别, 年龄, 文化程度, 住址, 电话
	顾客购物信息	顾客号, 商品号, 售价, 购买日期, 购买量

1.2 数据仓库

- 数据仓库的概念和特点（续）
  - 特点
    - 数据的集成性
      - 数据仓库中存储的数据是从原来分散在各个子系统中的数据提取出来的，经过处理后得到的。

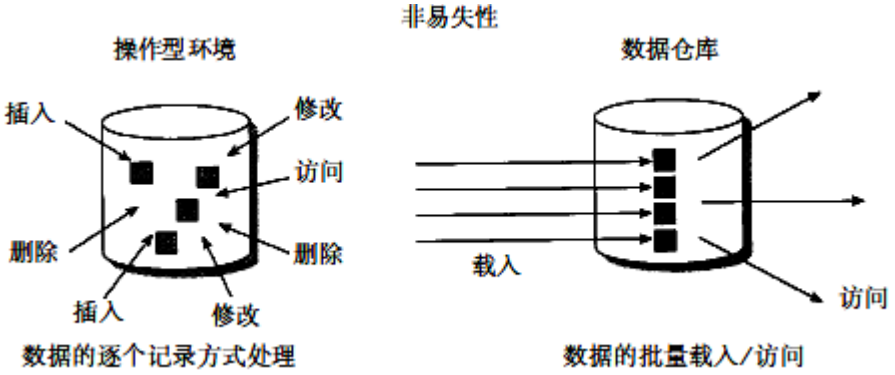
1.2 数据仓库

- **数据的集成性**



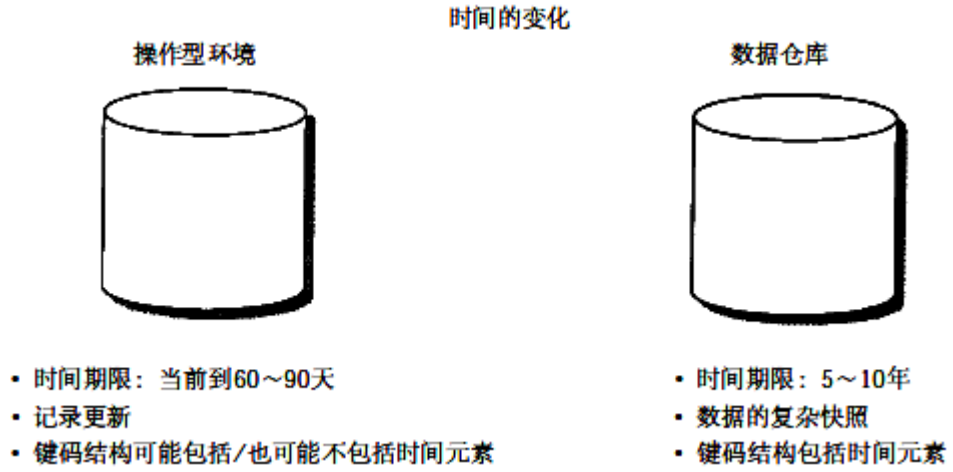
1.2 数据仓库

- 数据仓库的概念和特点（续）
  - 特点
    - 数据不可修改性
      - 数据仓库中的数据是不可更新的，只能通过分析工具进行查询、分析。



1.2 数据仓库

- 数据仓库的概念和特点（续）
  - 特点
    - 数据与时间相关
      - 数据随时间变化而定期地被更新



1.2 数据仓库

❖数据仓库与传统数据库的比较

传统数据库（事务性）数据	数据仓库（决策支持）数据
面向应用：数据服务于某个特定的商务过程或功能（OLTP）。	面向主题：数据服务于某个特定的商务主题，例如客户信息等。它是非规范化数据（OLAP）。
细节数据，例如包含了每笔交易的数据。	对源数据进行摘要，或经过复杂的统计计算。例如一个月中交易收入和支出的总和。
结构通常不变	结构是动态的，可根据需要增减。
易变性（数据可改变）	非易变（数据一旦插入就不能改变）。
事务驱动	分析驱动。
一般按记录存取，所以每个特定过程只操作少量数据。	一般以记录集存取，所以一个过程能处理大批数据，例如从过去几年数据中发现趋势。
反映当前情况。	反映历史情况。
通常只作为一个整体管理。	可以分区管理。
系统性能至关重要，因为可能有大量用户同时访问。	对性能要求较低，同时访问的用户较少。

1.2数据仓库

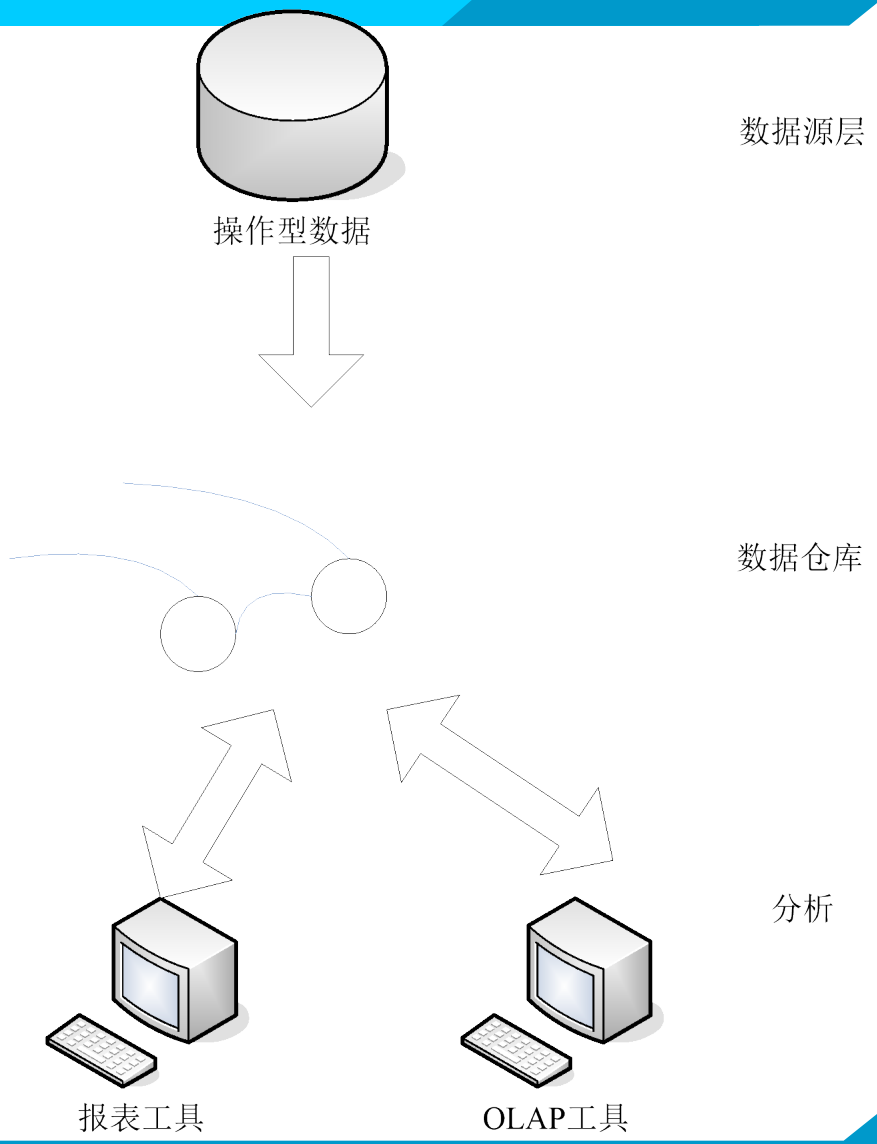
- **数据仓库带来的好处**
  - **提供决策支持。**
  - **应用于证券、银行、保险、移动通讯、商品销售、其它等行业。**

1.3数据仓库体系结构

- 数据仓库系统的特点
  - 分离性
  - 可扩展性
  - 安全性
  - 可管理性
- 体系结构分类
  - 面向结构的类型
    - 单层体系结构，两层体系结构，三层体系结构
  - 面向应用层次的类型
    - 独立数据集市，星型结构，联盟体系结构

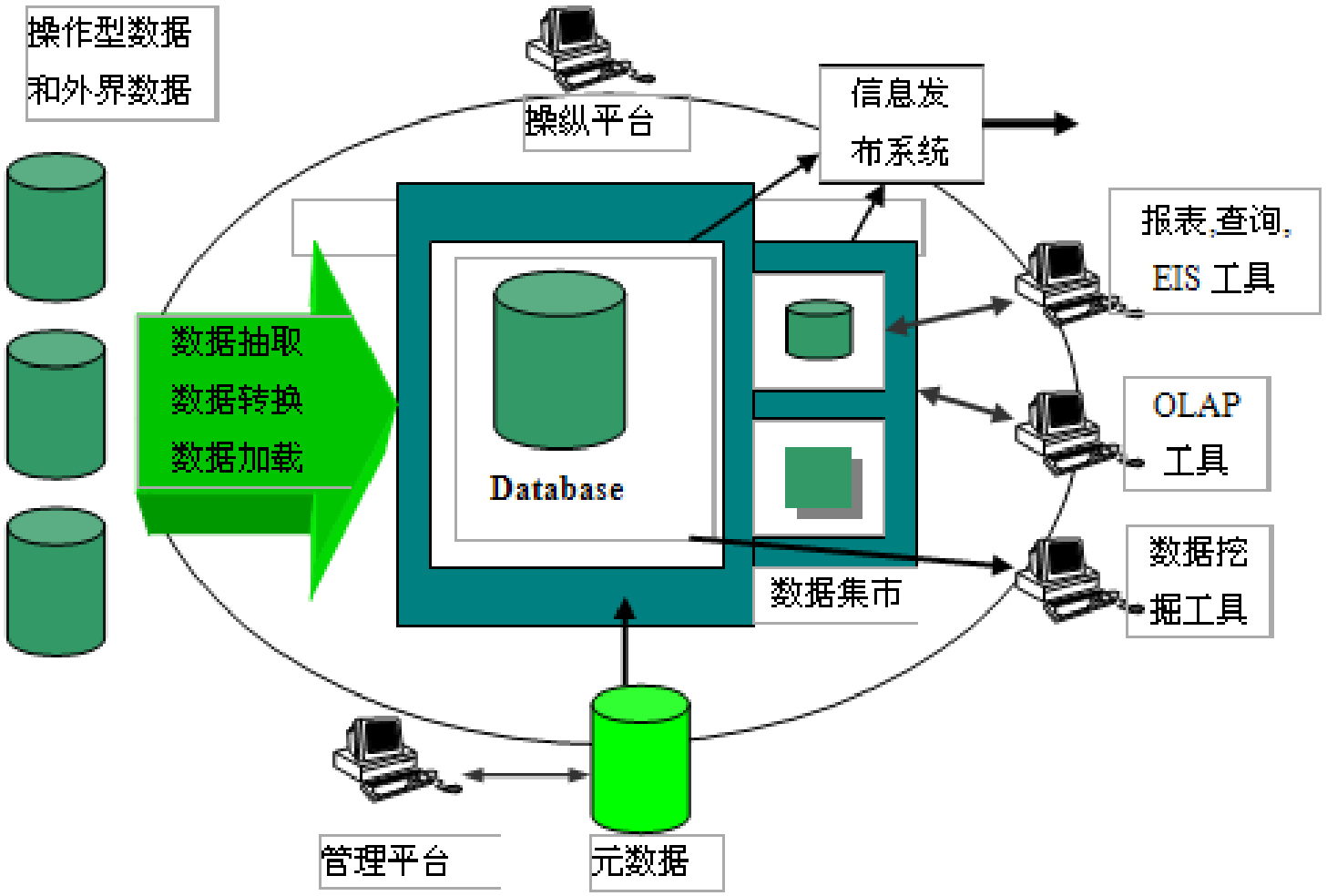
1.3数据仓库体系结构

- 面向结构的类型
  - 单层体系结构



1.3 数据仓库的体系结构

两层体系结构



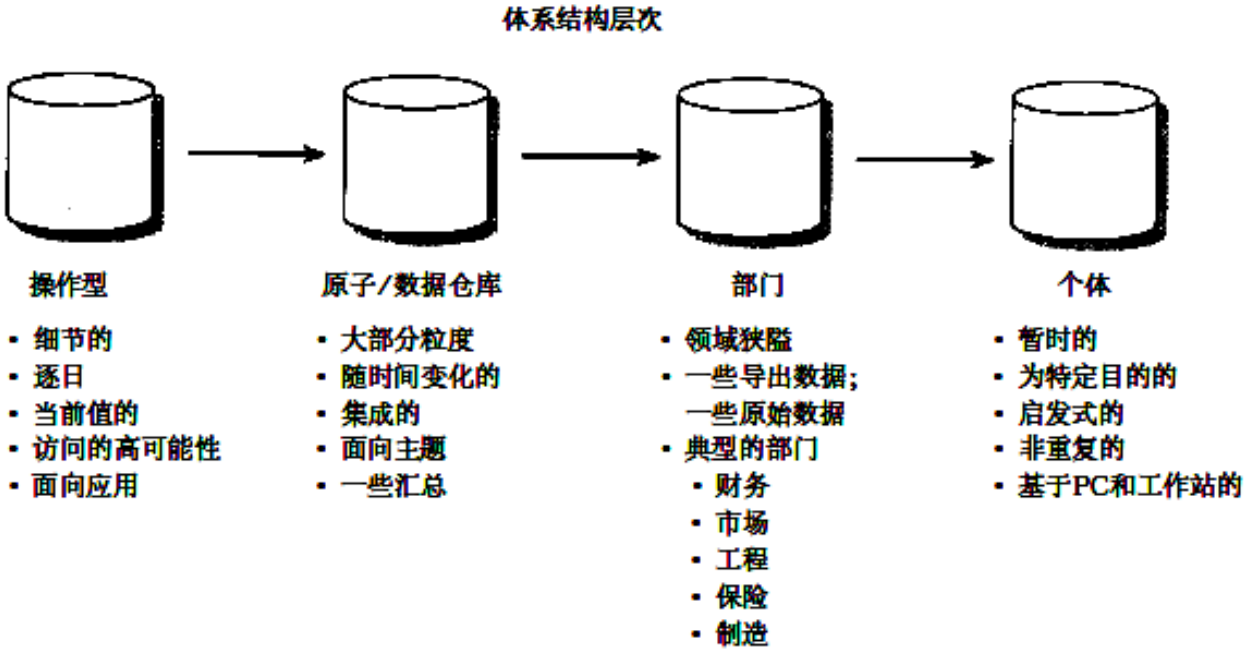
数据仓库体系结构示意图

1.3数据仓库体系结构

- 数据流
  - 数据源层
    - 关系数据库、或其它系统数据库
  - 数据准备
    - 提取、转换、加载（ETL）
  - 数据仓库层
    - 数据仓库、数据集市
  - 分析
    - 报表、信息分析、OLAP、数据挖掘

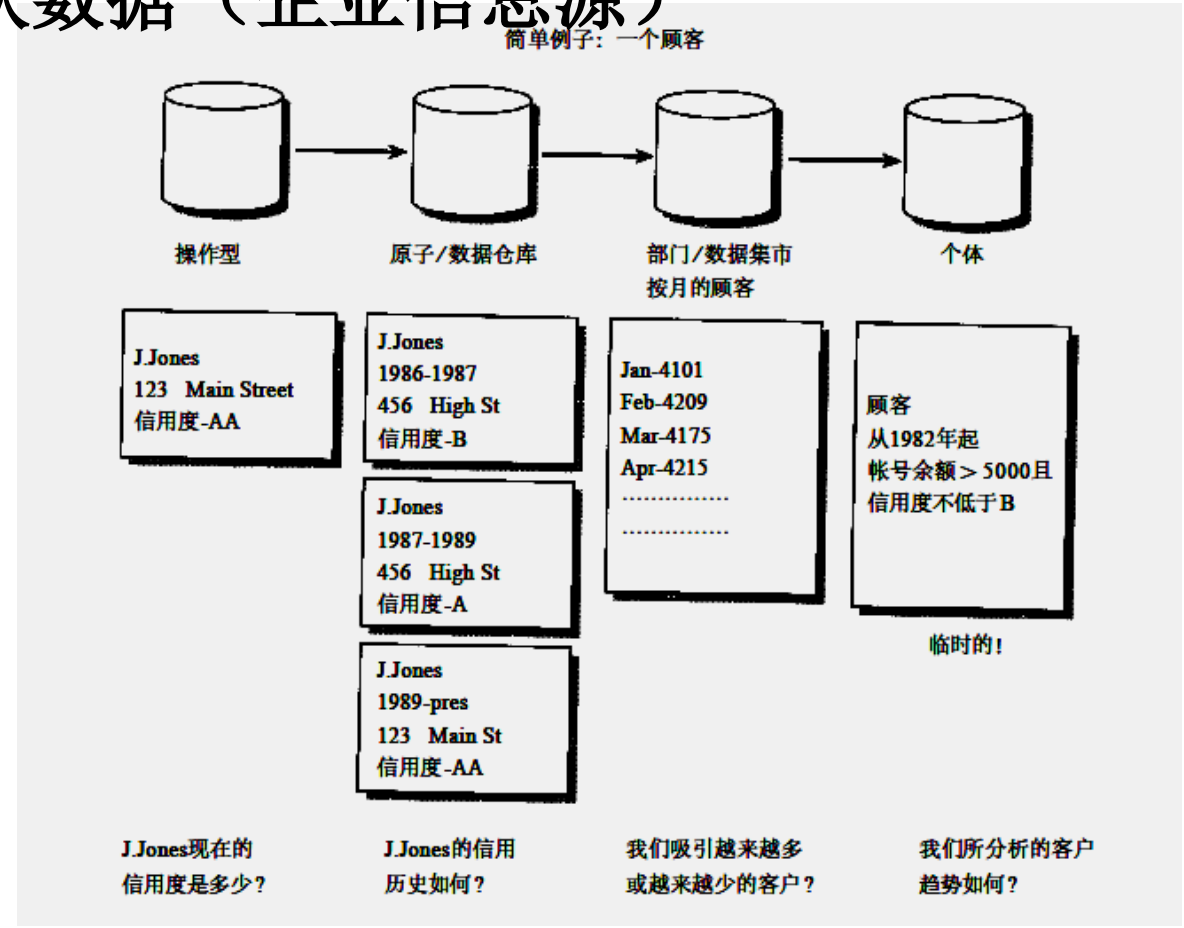
1.3数据仓库体系结构

- 层次数据（企业信息源）



1.3数据仓库体系结构

- 层次数据（企业信息源）



1.3数据仓库体系结构

- 数据集市（Data Marts）
  - 定义
    - 为了特定的应用目的或应用范围，而从数据仓库中独立出来的一部分数据，也称部门数据或主题数据。
    - 如：财务部门的数据集市
  - 与数据仓库的关系
    - 数据仓库是基于整个企业的数据模型建立的，它面向企业范围内的主题。而数据集市是按照某一特定部门的数据模型建立的。

1.3数据仓库体系结构

- 数据集市（Data Marts）（续）
  - 特性
    - 规模小
    - 特定的应用
    - 面向部门
    - 由业务部门定义，设计和开发
    - 由业务部门管理和维护
    - 快速实现
    - 购买较便宜
    - 投资快速回收
    - 更详细的、预先存在的数据仓库的摘要子集
    - 可升级到完整的数据仓库

1.3数据仓库体系结构

- 数据集市（Data Marts）（续）