数据仓库概述


马 帅

北京航空航天大學
BEIHANG UNIVERSITY

提纲

- DSS自然演化的体系结构
- 从数据库到 数据仓库
- 数据仓库定义
- 数据仓库中的数据组织

本次课程的用例：商场管理信息系统

- 进、销、存为主线，会员制
- 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）
- 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）

本次课程的用例：商场管理信息系统

- 库存子系统
  - 出库领料单 (出库领料单号，领料人，商品号，数量，日期)
  - 进料入库单 (进料入库单号，订单号，进料人，收料人，日期）
  - 库存台帐（商品号，库房号，库存量，日期）
- 人事子系统
  - 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，部门经理，电话）

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
  - 一笔订购（一张订单输入+订单细则）
  - 一笔销售（一张销售单）
  - 一次进料（一张进料单）
  - 一次出料（一张出料单）

强调多用户并发环境，数据的一致性、完整性

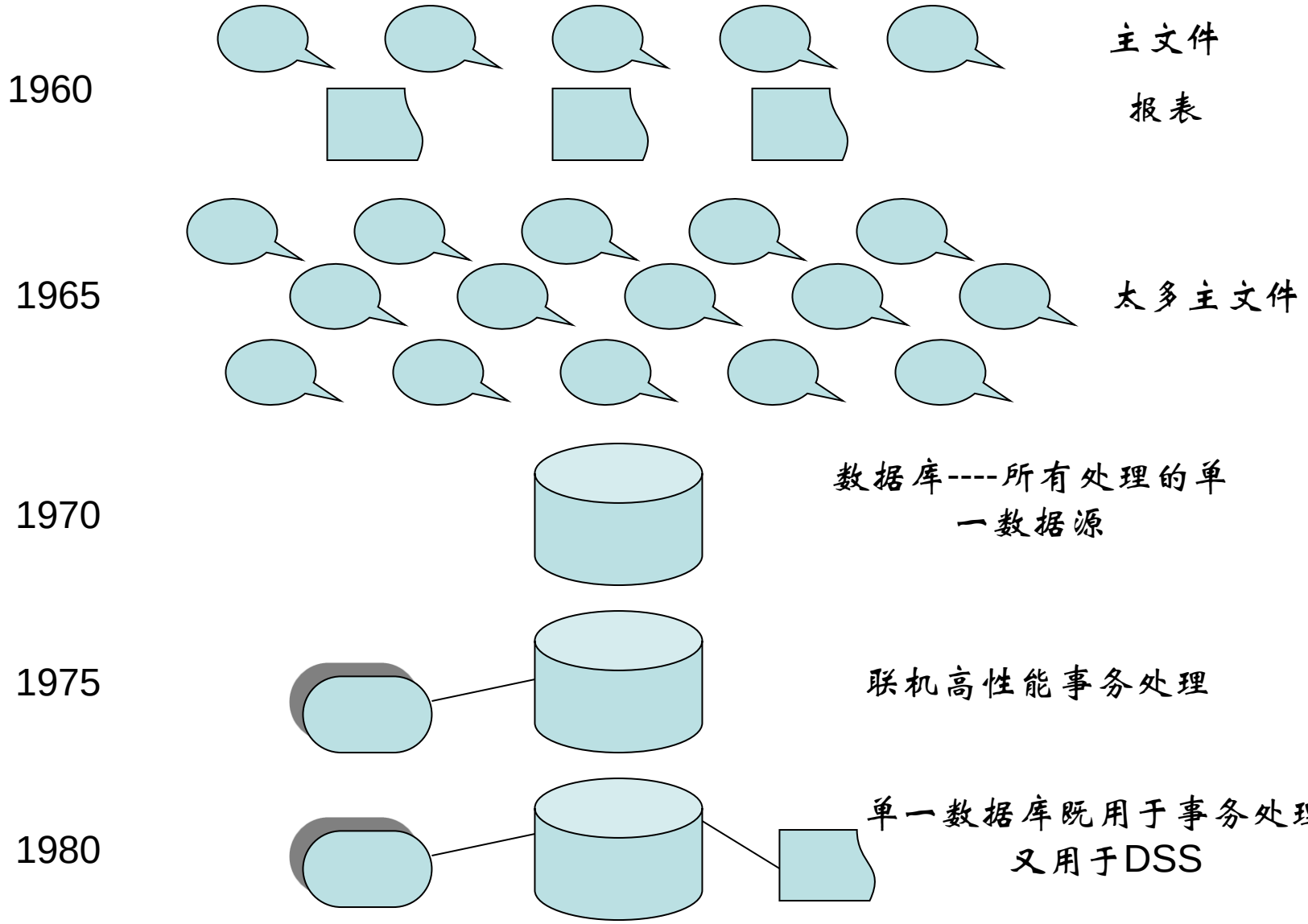
分析处理的需求

- 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）

分析处理的需求

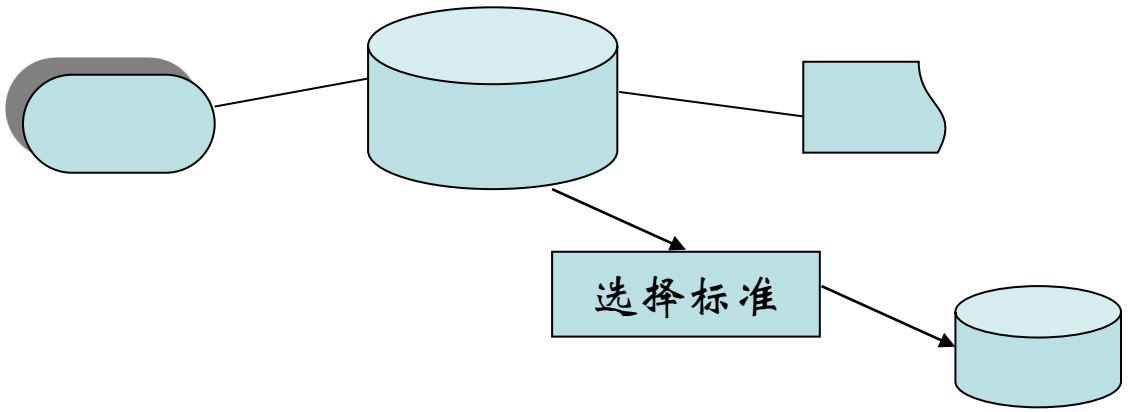
- 例2：某种商品今年的销售情况与以往相比，有怎样的变化？每年的第一季度商品销售在各类商品上的分布情况怎样？
- 要求：
  - 多个子系统中的数据（数据集成）
  - 历史数据
  - 汇总、综合的数据

DSS早期演化阶段


1960
主文件
报表
1965
太多主文件
1970
数据库----所有处理的单一数据源
1975
联机高性能事务处理
1980
单一数据库既用于事务处理，又用于DSS

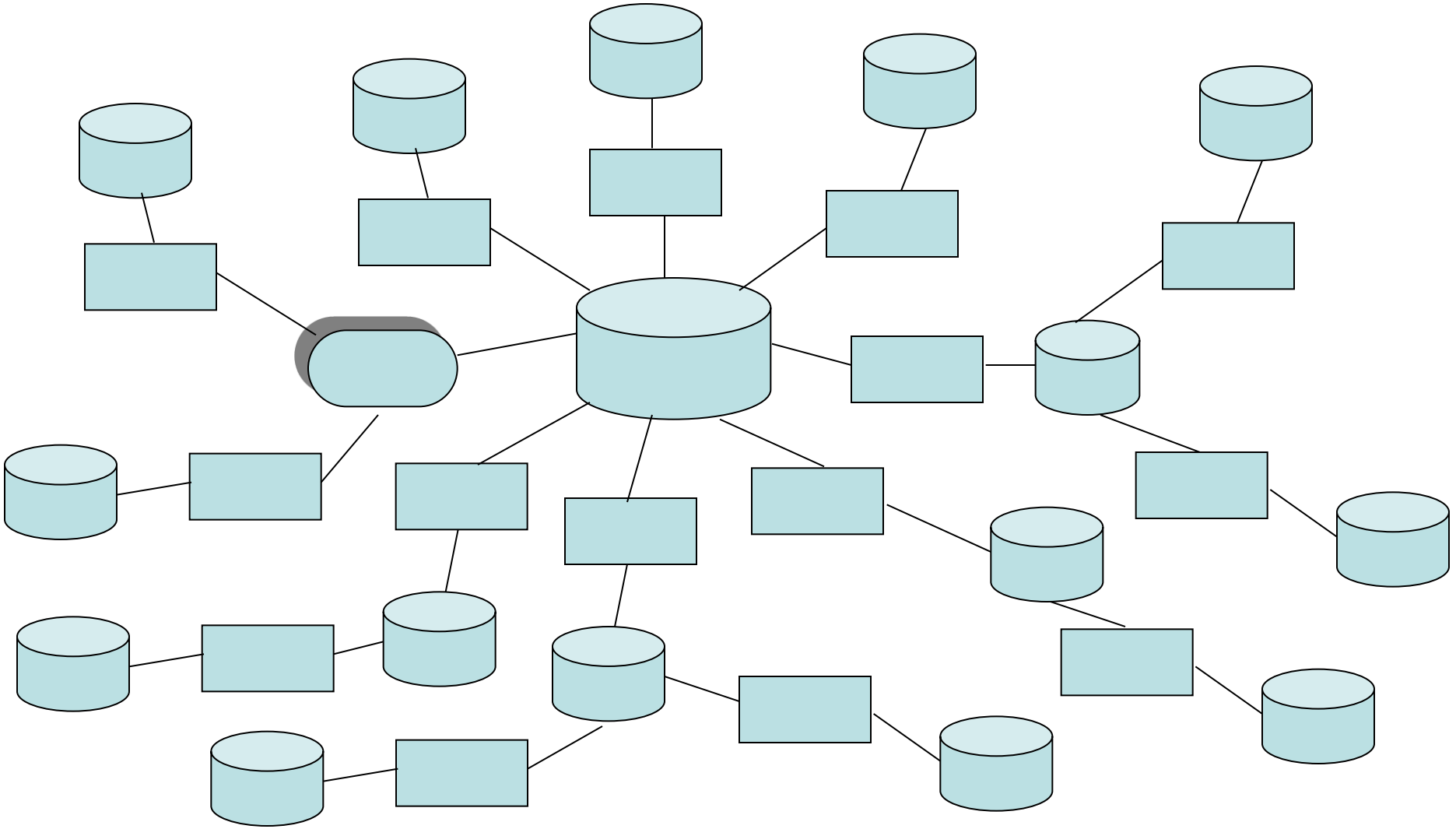
自然演化的体系结构

- 抽取程序
  - 搜索整个文件和数据库，使用某些标准选取合乎限制的数据，并把数据传到其他文件或数据库中
  - 优点
    - ✓ 将数据从事务处理应用中转移出来，在进行数据分析时不会与事务处理发生冲突
    - ✓ 当将数据从事务处理应用中抽取出来之后，数据的控制方式发生了转变，最终用户可以拥有抽取出来的数据



抽取程序

自然演化的体系结构



蜘蛛网

自然演化的体系结构

- 自然演化的体系结构中存在的问题
  - 数据缺乏可信性
  - 生产率
  - 数据转换为信息的不可行性

数据可信性

- 数据没有同一时间基准
  - 例如：一个银行的两个部门对同一业务提交报告
    - ✓ 部门A，于星期天傍晚提交，业务增长了10%
    - ✓ 部门B，于星期三下午提交，业务增长了15%
- 算法不同

  部门A使用的是所有类别的帐户

  部门B使用的是所有大帐户
- 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，然后又从此再次进行抽取，从数据进入系统到提供分析往往经过8、9次的抽取（误差累积）

数据可信性

- 外部数据问题

  一位分析员把华尔街杂志上的数据带进系统

  另一位将商业周刊的数据进入系统

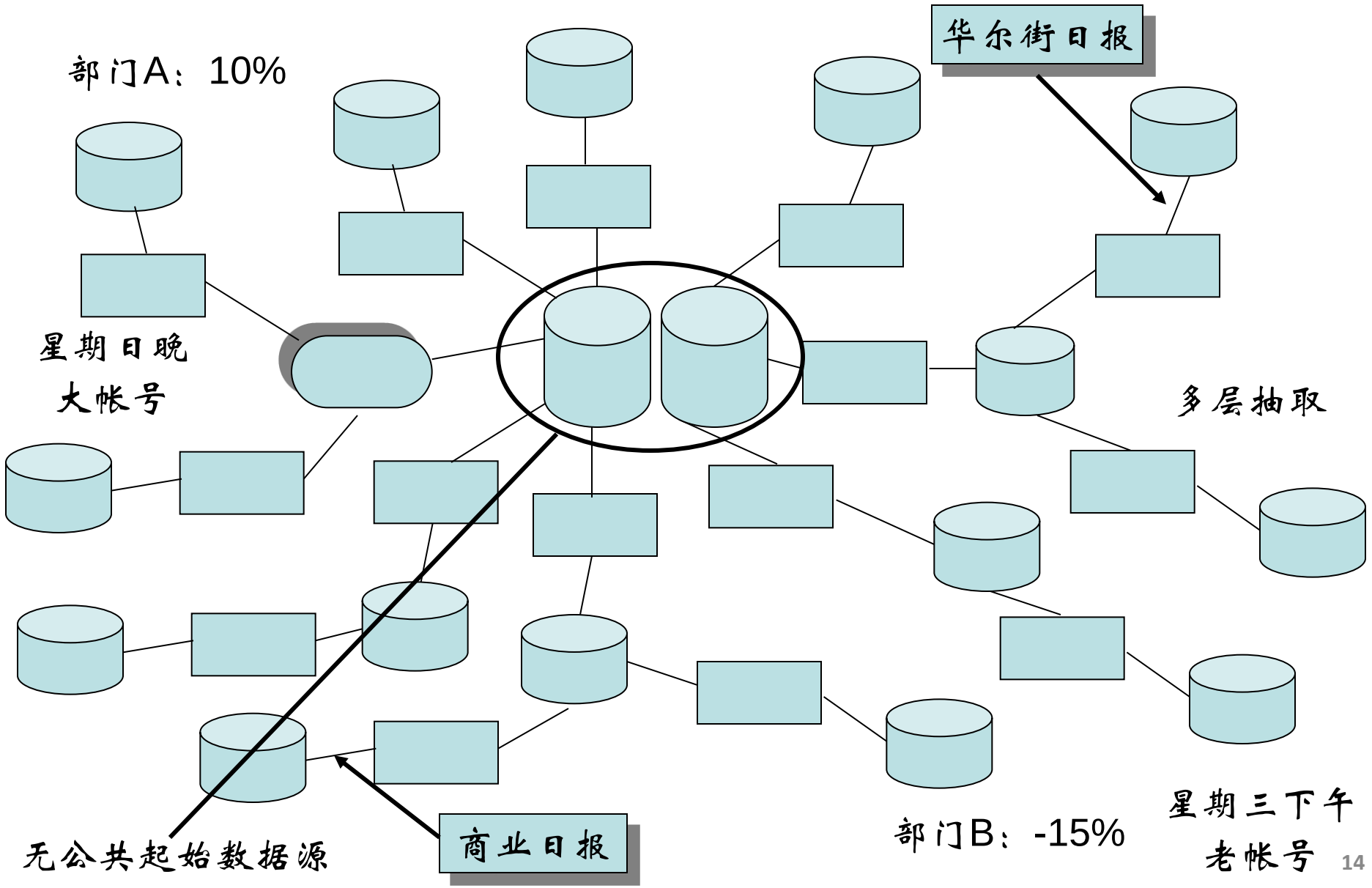
  数据一旦进入系统，往往已失去“身份”，并且一位分析员也不知道另一位分析员所输入的数据

- 无起始公共数据源

  部门A最初来源于文件XYZ

  部门B最初来源于数据库ABC

数据可信性



生产率

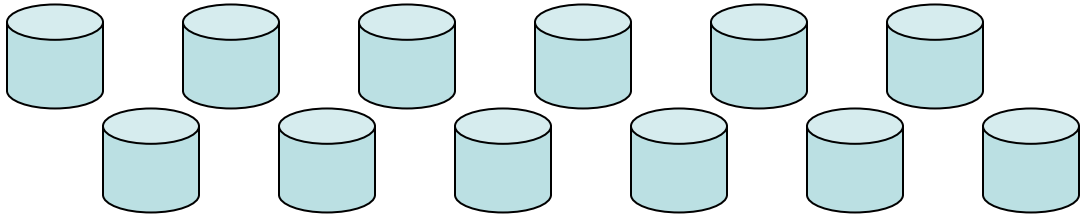
- 为了生成一个企业报表，必须经过
  - 获得源数据
    - ✓ 很多文件和数据库
  - 定位和分析数据
    - ✓ 需要分析很多文件和数据布局
    - ✓ 数据不一致造成很难准确定位和分析
      - 同名不同义：如weight字段在A表中表示人的体重，在B表中表示汽车的重量
      - 同义不同名：如在A表中的balance字段，在b表中为bal
      - 结构不同：同一字段在不同的应用中数据类型不同。如Sex字段在A表中为“M/F”，在B表中为“0/1”

生产率

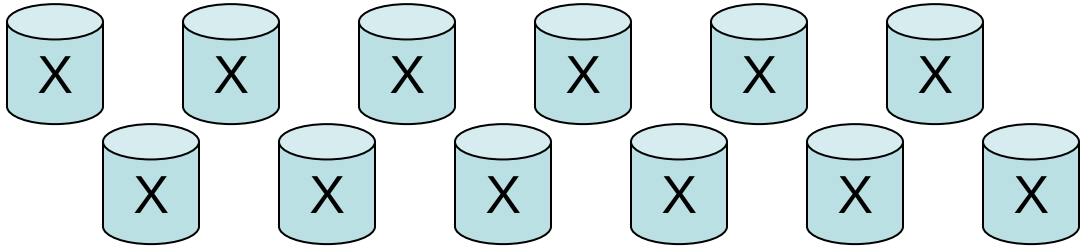
- 把数据加工成报表
  - ✓ 要写许多程序，每个程序必须客户化（与客户环境有关）
  - ✓ 程序会涉及公司具有的各种技术
  - ✓ 由于定位数据困难，检索所要的数据是一件很麻烦的事
- 完成任务需要很长时间
  - ✓ 定位数据 + 获得数据 + 集成报告，完成任务所需时间较长
  - ✓ 每份报告各自需求不同，前面报表不会为后继报表提供什么帮助，因此每份报告所需要的时间都很长，

生产率

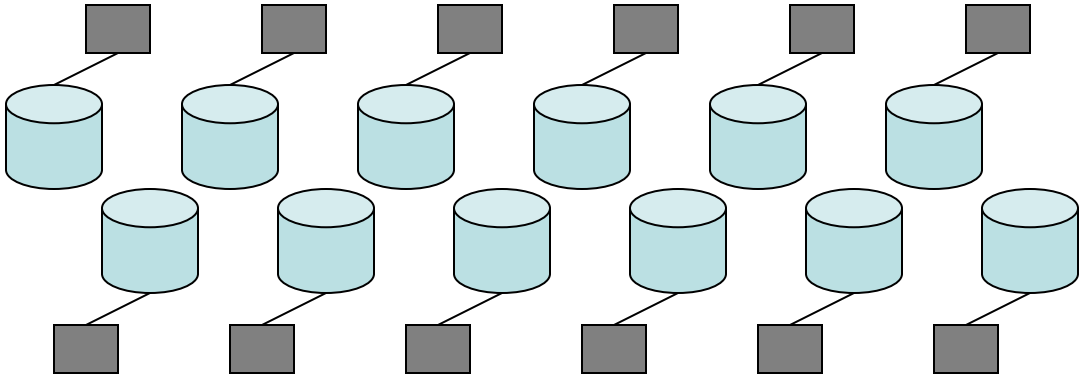
根据全部企业数据生成报表



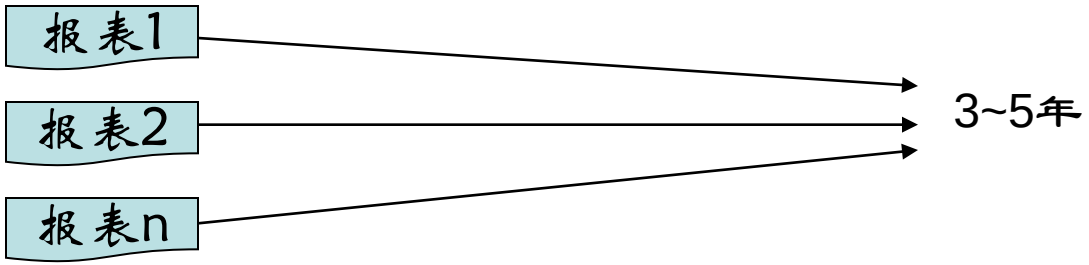
定位数据需要浏览大量文件



得到数据：抽取程序各不相同，每个都是定制的



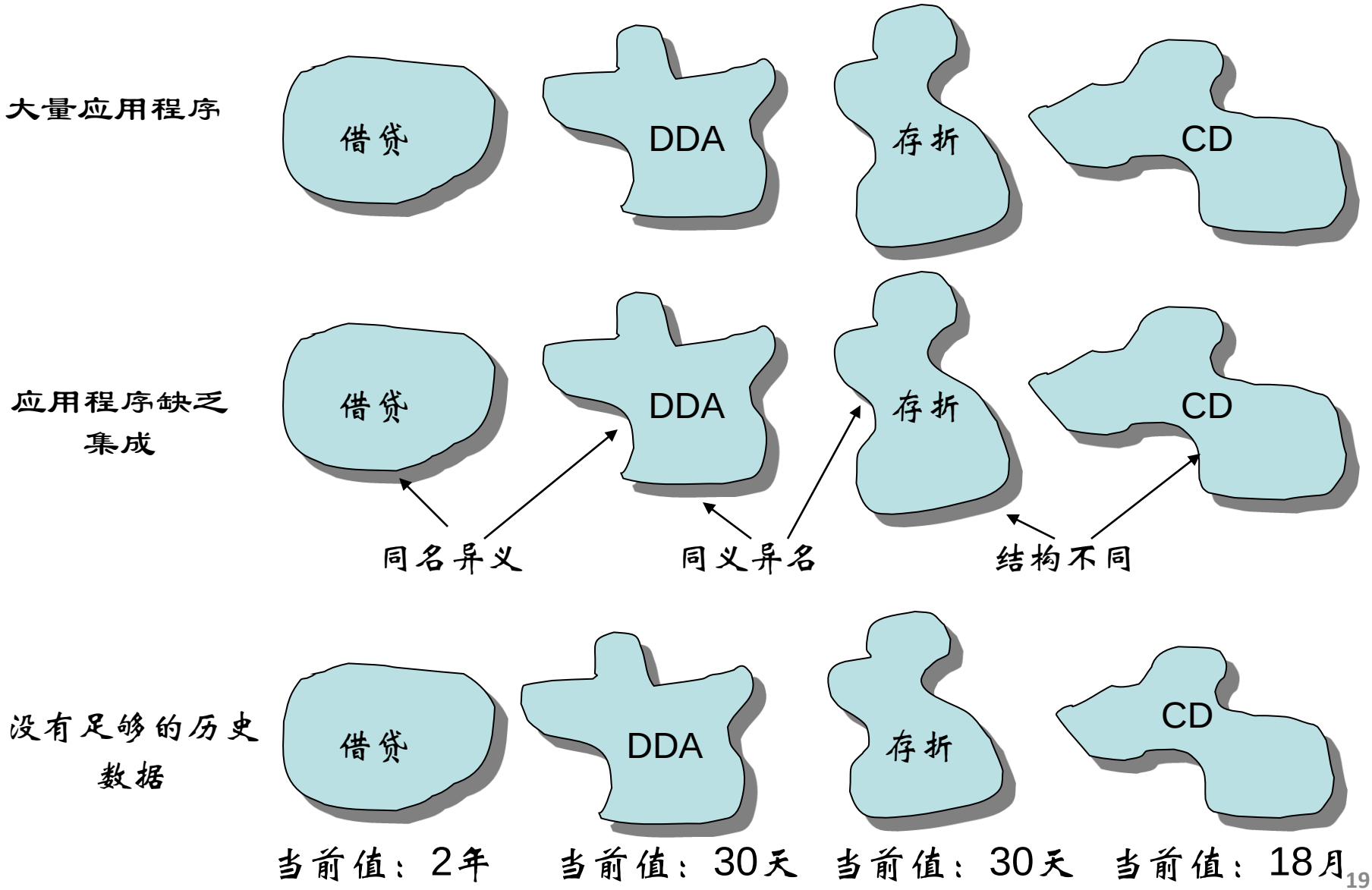
利用率低：编写第一张报表时对后继报表的需求不清楚



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（Savings）；贷款（Loan）；活期存款记帐（DDA）；信托（Trust）
  - 这些应用并未集成
    - ✓ 数据不一致问题
    - ✓ 外部数据和非结构化数据
  - 没有足够的历史数据
    - ✓ 例如：贷款部门有二年的数据；银行存折处理，拥有一年的数据，DDA只有60天的数据；现金交易处理具有18个月的数据

从数据到信息


大量应用程序
借贷
DDA
存折
CD
应用程序缺乏集成
借贷
DDA
存折
CD
同名异义
同义异名
结构不同
没有足够的历史数据
借贷
DDA
存折
CD
当前值：2年
当前值：30天
当前值：30天
当前值：18月

从数据库到数据仓库

操作型数据	分析型数据
细节的	综合的，或提炼的
在存取瞬间是准确的	代表过去的数据
可更新	不更新
操作需求事先可知道	操作需求事先不知道
生命周期符合SDLC	完全不同的生命周期
对性能要求高	对性能要求宽松
一个时刻操作一单元	一个时刻操作一集合
事务驱动	分析驱动
面向应用	面向分析
一次操作数据量小	一次操作数据量大
支持日常操作	支持管理需求

两种数据的区别

数据仓库的定义

数据仓库是一个**面向主题的**（Subject Oriented），**集成的**（Integrated），**相对稳定的**（Nonvolatile），**反映历史变化的**（time Variant）数据集合。用于支持管理决策

数据仓库的特点：面向主题

面向应用（操作）

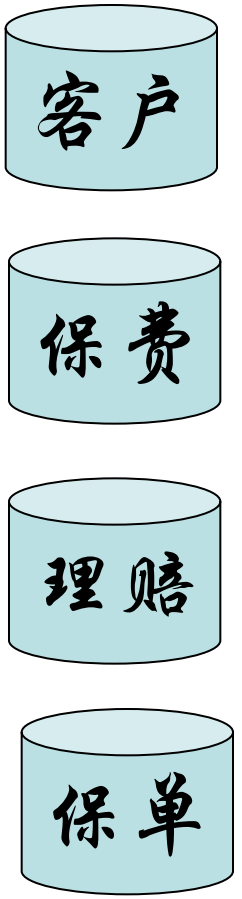
机动车险

寿险

健康险

意外险

面向主题（分析）



数据仓库的特点：面向主题

- 主题
  - 企业中某一宏观分析领域所涉及的分析对象
  - 在较高层次上将企业信息系统中的数据综合、归类并进行分析利用的抽象
- 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，能完整、统一地刻画各个分析对象所涉及的企业的各项数据，以及数据之间的联系
- 数据组织
  - 数据库：面向应用
  - 数据仓库：面向主题

数据仓库的特点：面向主题

- 面向应用的数据组织
  - 面向应用进行数据组织是对企业中相关的组织、部门等进行详细调查，收集数据库的基础数据及其处理的过程

    “数据”→“数据字典”，“处理”→“数据流图”
  - 面向应用的数据组织应反映一个企业内数据的动态特征，即要便于表达企业各部门内的数据流动情况及部门间的数据输入输出关系；应按实际业务处理流程来组织数据，以便进行联机事务处理

数据仓库的特点：面向主题

- 面向应用的数据组织所生成的数据库模式与实际业务处理流程中涉及的单据或文档有良好的对应关系
- 面向应用的数据组织没有体现数据库的“数据与处理的分离”本质思想
  - ✓ 数据库的建设偏重于对联机事务处理的支持
  - ✓ 数据的应用逻辑与数据本身有一定程度的捆绑
  - ✓ 原本描述同一实体的数据由于与不同的应用逻辑捆绑而变得不统一
  - ✓ 原本描述同一实体的数据分散在不同的数据库模式中
- 面向应用的数据组织的抽象程度不够高，没有完全实现数据与应用的分离，但能很好地支持OLTP

数据仓库的特点：面向主题

- 示例：以商品采购为例
  - OLTP
    - ✓ 描述一笔采购业务
    - ✓ 模式：订单、订单细则、供应商
  - OLAP
    - ✓ 关心采购渠道
    - ✓ 按“供应商”重新组织数据
    - ✓ 供应商基本（固有）信息：供应商号，供应商品、地址、电话
    - ✓ 供应商品信息：供应商号、商品号、供应价、供应量、日期

数据仓库的特点：面向主题

- 例：商场销售系统所三个主题
  - 商品
    - ✓ 商品基本信息
      商品号，商品名，类别，颜色等
    - ✓ 商品采购信息
      商品号，供应商号，供应价，供应日期，供应量等
    - ✓ 商品销售信息
      商品号，顾客号，售价，销售日期，销售量等
    - ✓ 商品库存信息
      商品号，库房号，库存量，日期等；

数据仓库的特点：面向主题

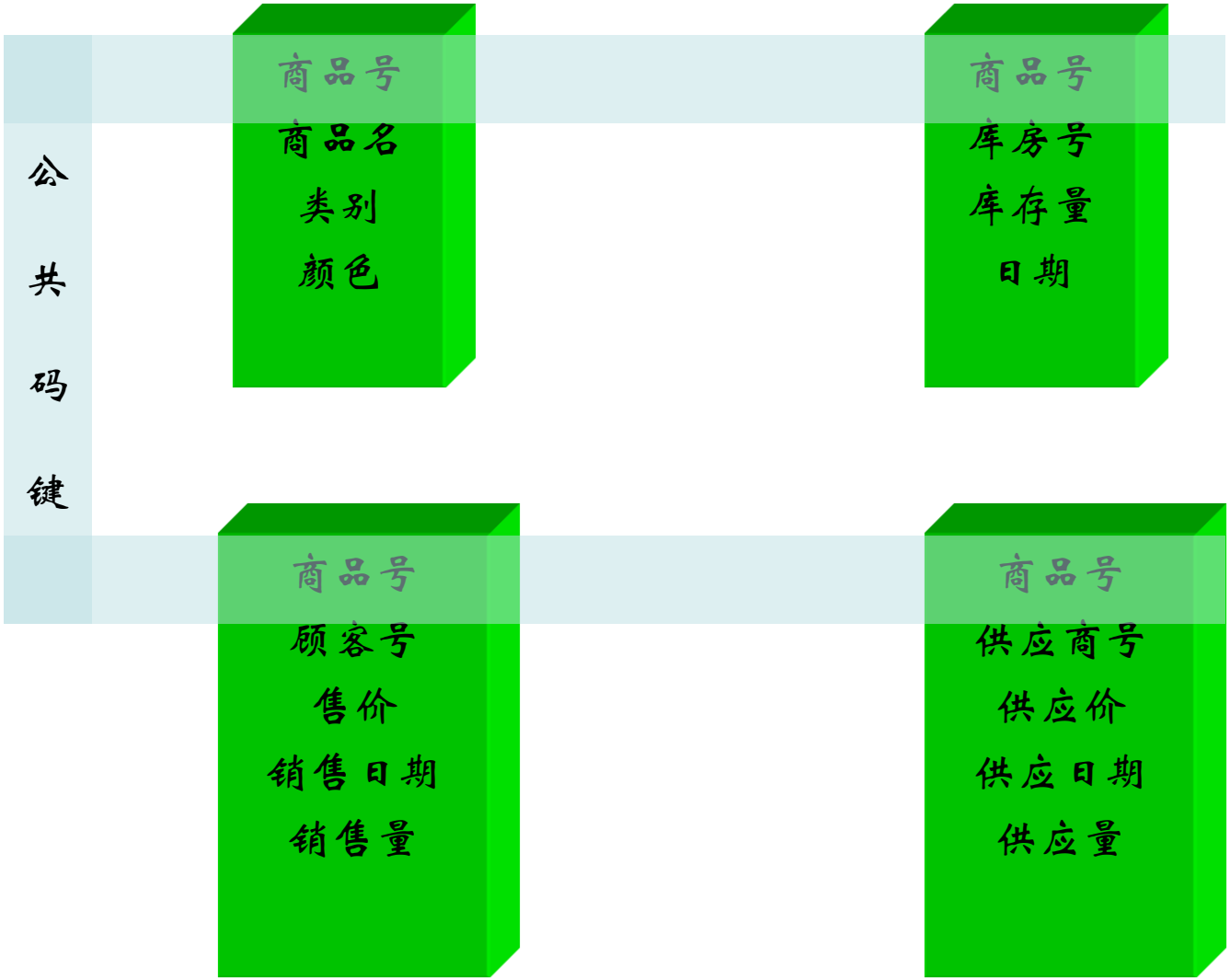
- 供应商
  - ✓ 供应商基本信息
    供应商号, 供应商名, 地址, 电话, 供应商类型等
  - ✓ 供应商品信息
    供应商号，商品号，供应价，供应日期，供应量等
- 顾客
  - ✓ 顾客基本信息
    顾客号, 顾客名, 性别, 年龄, 文化程度, 住址，电话等
  - ✓ 顾客购物信息
    顾客号，商品号，售价，购买日期，购买量等

数据仓库的特点：面向主题

- 面向主题数据组织方式的实现
  - 对应多个表，通过公共码键把各个表统一联系起来，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）

数据仓库的特点：面向主题



数据仓库的特点：面向主题

主题：　　商品　　公共码键：　商品号

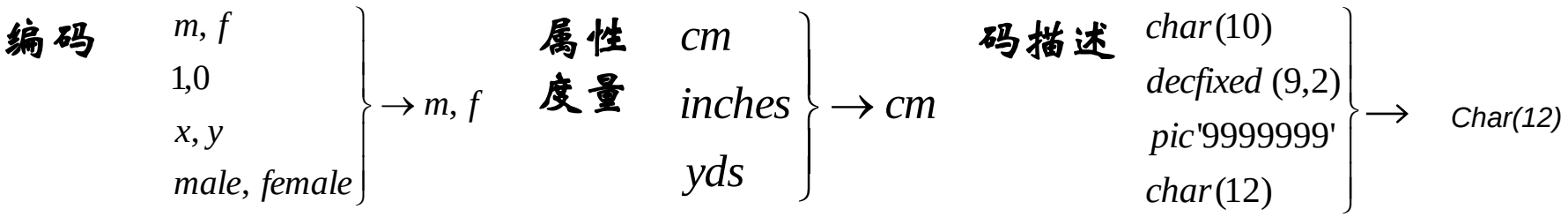
商品表(商品号，商品名，类型,颜色 …)　　/* 描述的是商品的固有信息　　*/
采购表1(商品号，供应商号，供应日期，供应价，供应数量，…)
　　/* 描述的是商品的采购细节信息 */
采购表2(商品号，时间段，采购总量 ,…)　　/* 某时间段内商品采购信息　　*/
…
采购表n(…, …)　　/* 时间段不等的采购综合表　　*/
销售表1(商品号，顾客号，销售日期，售价，销售量，…)
　　/* 描述的是商品的销售细节信息 */
销售表2(商品号，时间段，销售总量，…)　　/* 某时间段内商品销售信息　　*/
…
销售表n(…, …)　　/* 时间段不等的销售综合表　　*/
库存表1(商品号，库房号，库存量，日期，…)/* 描述的是商品的库存细节信息 */
库存表2(商品号，库房号，库存量，月份，…)/* 每月月底的商品库存信息　　*/
…
库存表n(…, …)　　/* 时点不同的商品库存信息　　*/

多个层次综合信息

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



- 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

- 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、完整性保护
  - 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：反映时间变化的

- 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
  - 按时间段进行综合
  - 隔一定的时间片进行抽样

数据仓库的特点：反映时间变化的

- 操作型数据与DW中的数据比较
  - 操作型
    - ✓ 60-90天数据
    - ✓ 含有当前值，能被更新
    - ✓ 码中不一定包括时间元素
  - **DW**
    - ✓ 5-10年数据
    - ✓ 一系列快照
    - ✓ 码中包括时间元素

数据仓库的数据组织

- 数据仓库的数据组织结构
- 粒度
- 样本数据
- 分割
- DW中的数据组织的实现方式
- 数据追加

数据仓库的数据组织结构

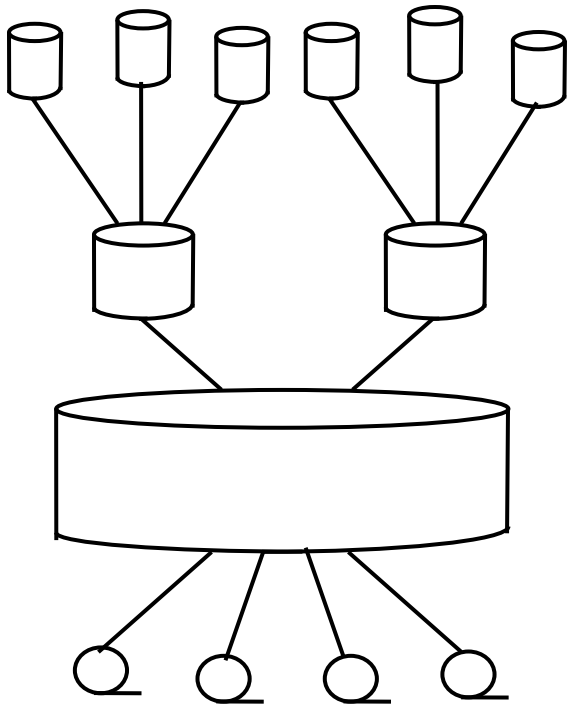
元数据

高度综合级→

轻度综合级→

当前细节级→

早期细节级→



1984－1989 每月销售数据

1984－1989 每周销售数据

1990－1991 销售细节数据

1984－1989 销售细节数据

源数据首先进入当前细节级

老化的数据进入早期细节级

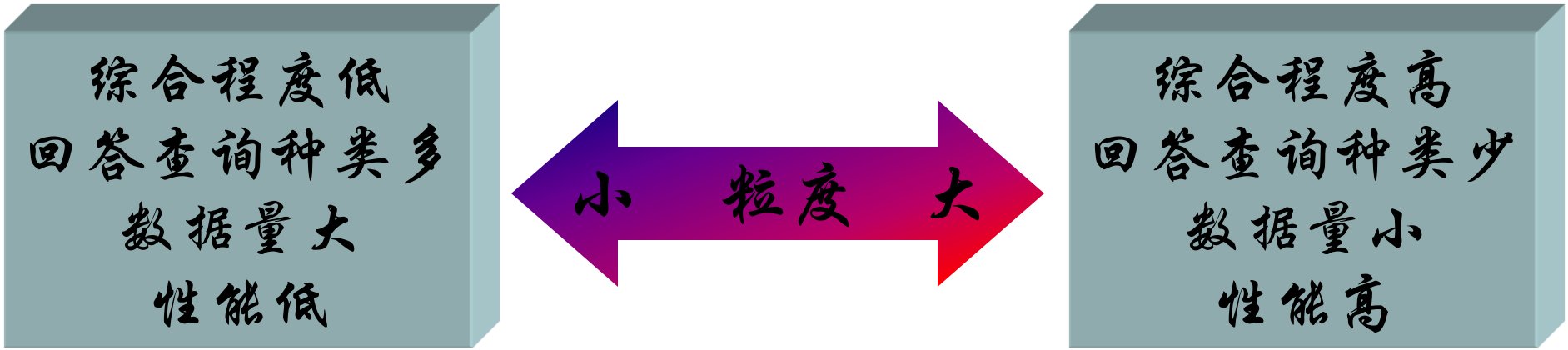
轻度综合级对应于数据集市

元数据（Metadata）

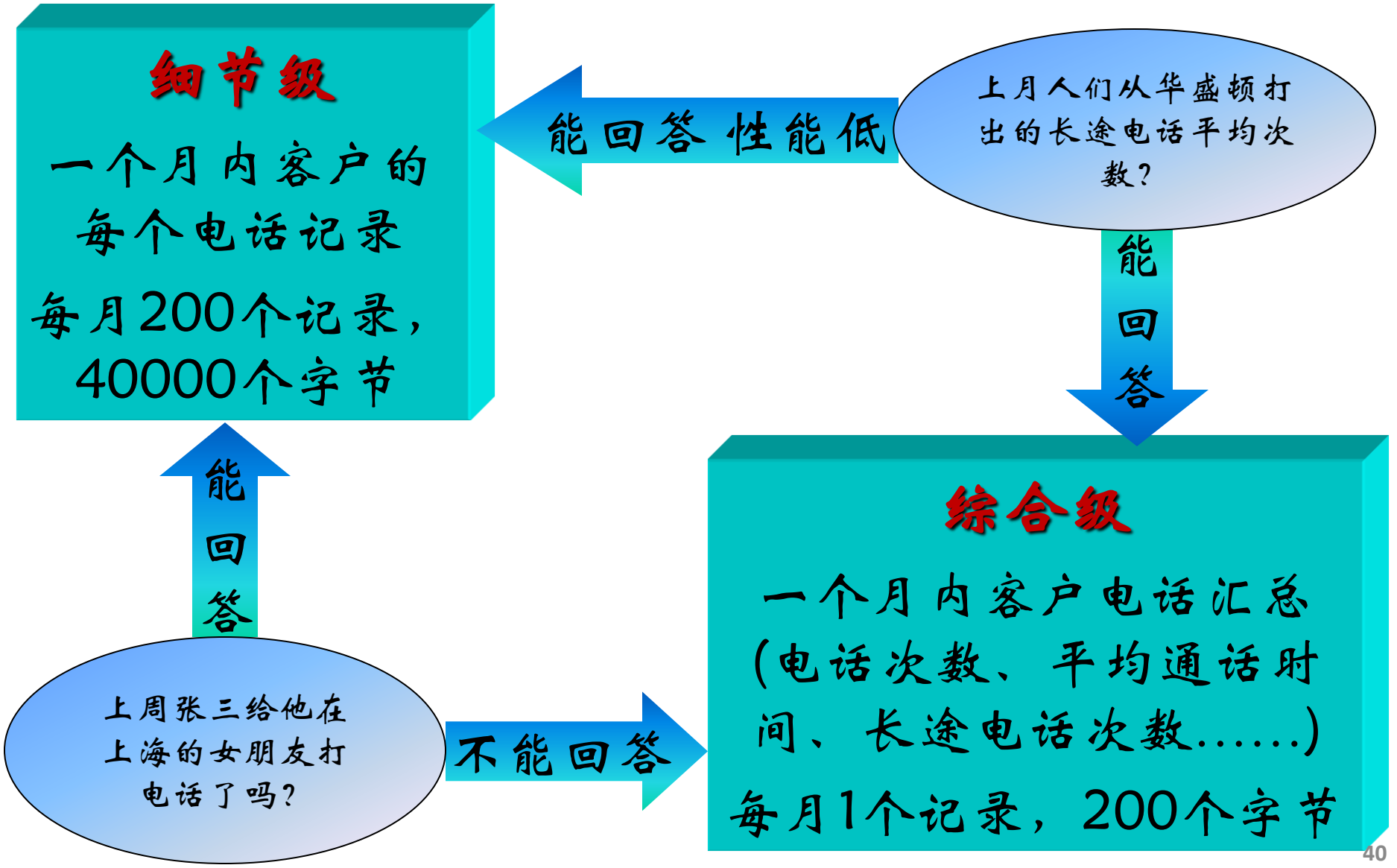
- 元数据
  - 关于数据的数据
  - 从DB → DW时， DB中的源数据描述及到DW中的映射
  - DSS元数据：数据从DW → OLAP（前端工具）之间的映射

数据粒度

- 粒度
  - 数据综合程度高低的一个度量
  - 粒度越小，越细节，综合程度越低，回答查询种类越多，数据量大，性能低

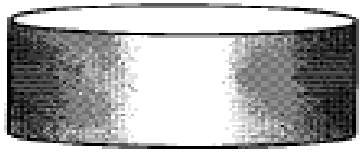


数据粒度


细节级
一个月内客户的每个电话记录
每月200个记录，40000个字节
能回答 性能低
上月人们从华盛顿打出的长途电话平均次数？
能回答
综合级
一个月内客户电话汇总(电话次数、平均通话时间、长途电话次数……)
每月1个记录，200个字节
不能回答
上周张三给他在上海的女朋友打电话了吗？
能回答

不同粒度级别比较

高细节级



例如：一个顾客一个月的每个电话的细节

低细节级

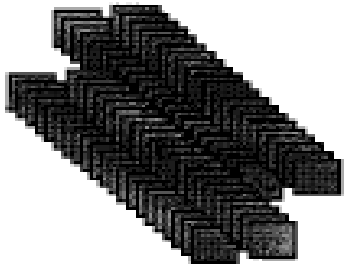
例如：一个顾客一个月的电话综合

"Cass Squire上星期是否给他在波士顿的女友打了电话？"

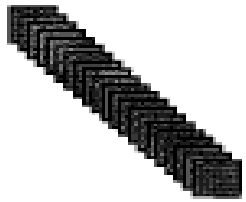
- 能回答，尽管需要一定数量的检索
- 根本就不能回答。细节已经消失

但寻找单个记录是个非常不常见的事件

"上个月，人们从华盛顿打出的长途电话平均有多少个？"



搜寻175 000 000个记录，进行45 000 000 次I/O

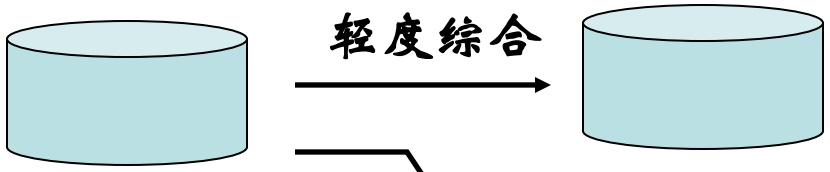


搜寻1 750 000个记录，进行 450 000次I/O

数据粒度

- 多重粒度
  - 适应多种分析处理需要
  - 大粒度数据放在快速设备如磁盘上
  - 小粒度数据放在慢速设备如磁带上
  - 由于大部分DSS分析都是基于一定程度的综合数据上的，因此上述布局可以可有效地处理绝大多数的请求，并回答任何能够回答的问题。这是最好的并且应作为默认的设计选择

数据粒度

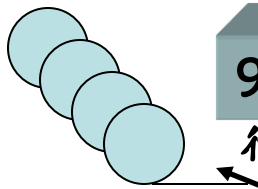


四月份
J Jones
电话次数—45 个
电话的平均长度—14 分钟
长途电话次数—18 个
接线员帮助呼叫次数—2 个
未接通电话次数—1 个

存储1个记录所需的字节数—225

J Jones
4月12日下午6:01～6:12
415-566-9982接线员帮助
4月12日下午6:15～6:16
415-334-8847长途
4月12日下午6:23～6:38
408-223-7745
……

每月一个顾客存储200 个记录
平均需要的字节数—45 000



95%甚至更多的DSS处理在此进行

真实档案

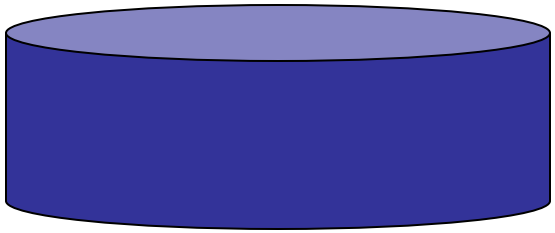
5% 或更少的DSS处理在此进行

- 耗费时间
- 复杂
- 费用高

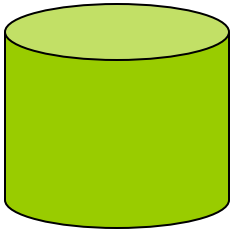
样本数据

- 样本数据
  - 特殊形式的数据粒度
  - 启发式分析
    - ✓ 运行程序 ≠分析结果 ≠修改程序 ≠再运行程序
    - ✓ 不要求准确的结果，只需要建立起分析模型或得到相对准确、能反映趋势的数据，以验证用户猜想，为下一步的策略确定方向或对当前分析程序作出相应调整
  - 样本数据库是以一定的采样率从细节档案数据或轻度综合数据中抽取的一个子集
  - 样本数据可以代替源数据进行模拟分析，以提高分析效率

样本数据



数据仓库



样本数据库

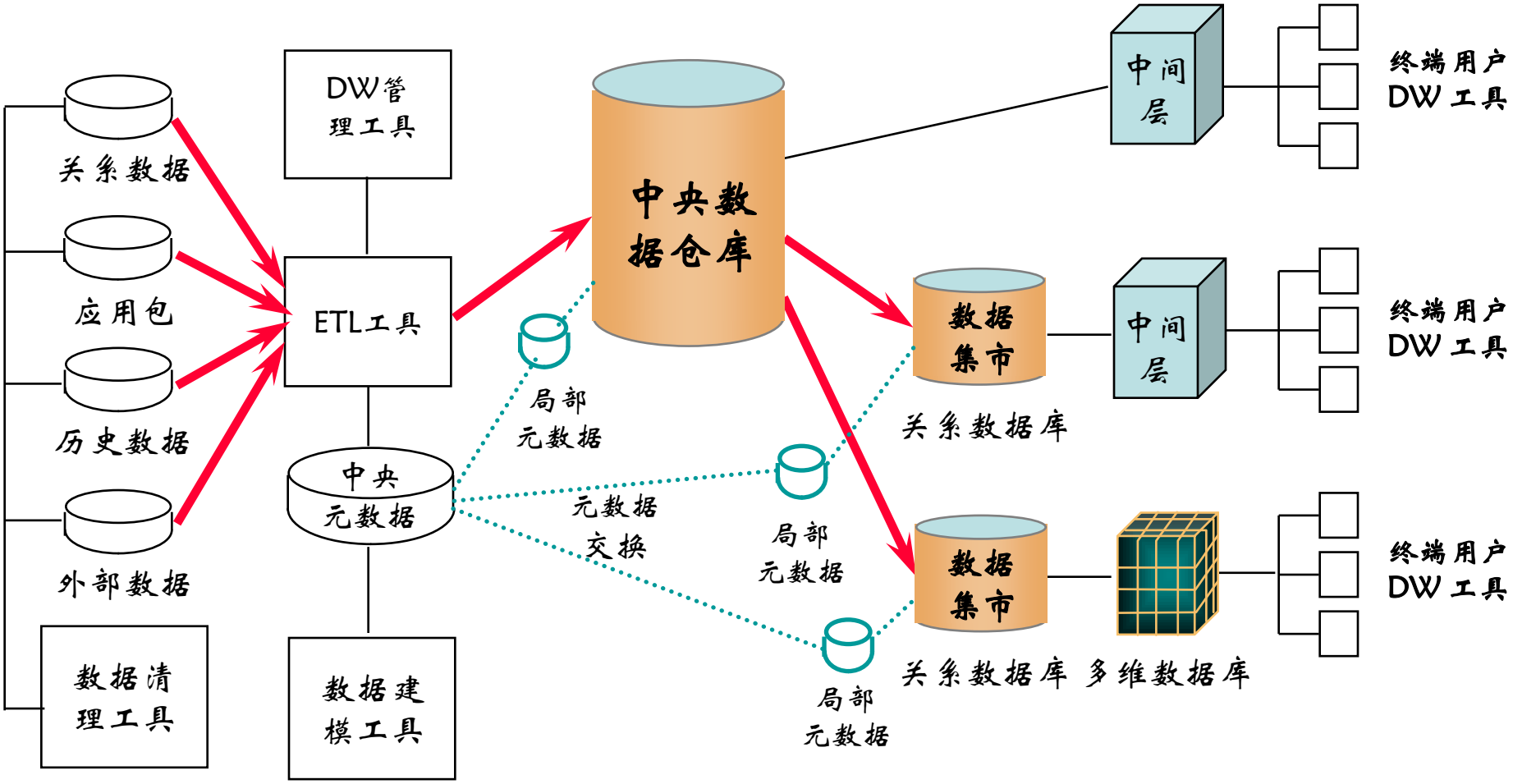
“所有的保险客户中，有多少是已婚且具有大学学历的 35 岁以上的男性？”

- 仓库中的一部分数据
- 用于查询的非常有效的形式
- 不能用于一般目的的分析—只能用于统计分析

样本数据库：一种改变数据粒度的方法

样本数据

- 样本数据的粒度以采样率的高低来划分，不同的采样粒度可有相同的综合级别
- 示例：商品库存表
  - ✓ 商品库存信息：每日库存信息表、每月（底）库存信息表。不同时点库存信息表，反映了不同的采样率（粒度），但可以是相同的综合级别。
- 示例：统计马路上男司机的比例
  - ✓ 当分析员遇到一个大文件，用325 000 000 条记录确定5 6 . 7 %上路的汽车司机是男人，而使用样本数据库，分析员用了25 000 个记录确定5 5 . 9 %上路的汽车司机是男人


关系数据
应用包
历史数据
外部数据
数据清理工具
DW管理工具
ETL工具
中央元数据
数据建模工具
中央数据仓库
局部元数据
元数据交换
局部元数据
局部元数据
中间层
中间层
数据集市
数据集市
关系数据库
关系数据库 多维数据库
终端用户DW工具
终端用户DW工具
终端用户DW工具
源数据库
数据抽取、转换、装载
中央数据仓库
数据集市
数据访问和分析工具

数据仓库的体系结构


谢谢