大数据库系统

第5章 数据仓库与数据挖掘

第5章 数据仓库与数据挖掘

主要内容

5.1 数据仓库简介

5.1.1 数据仓库的产生

5.1.2 数据仓库的定义

5.1.3 数据仓库的特征

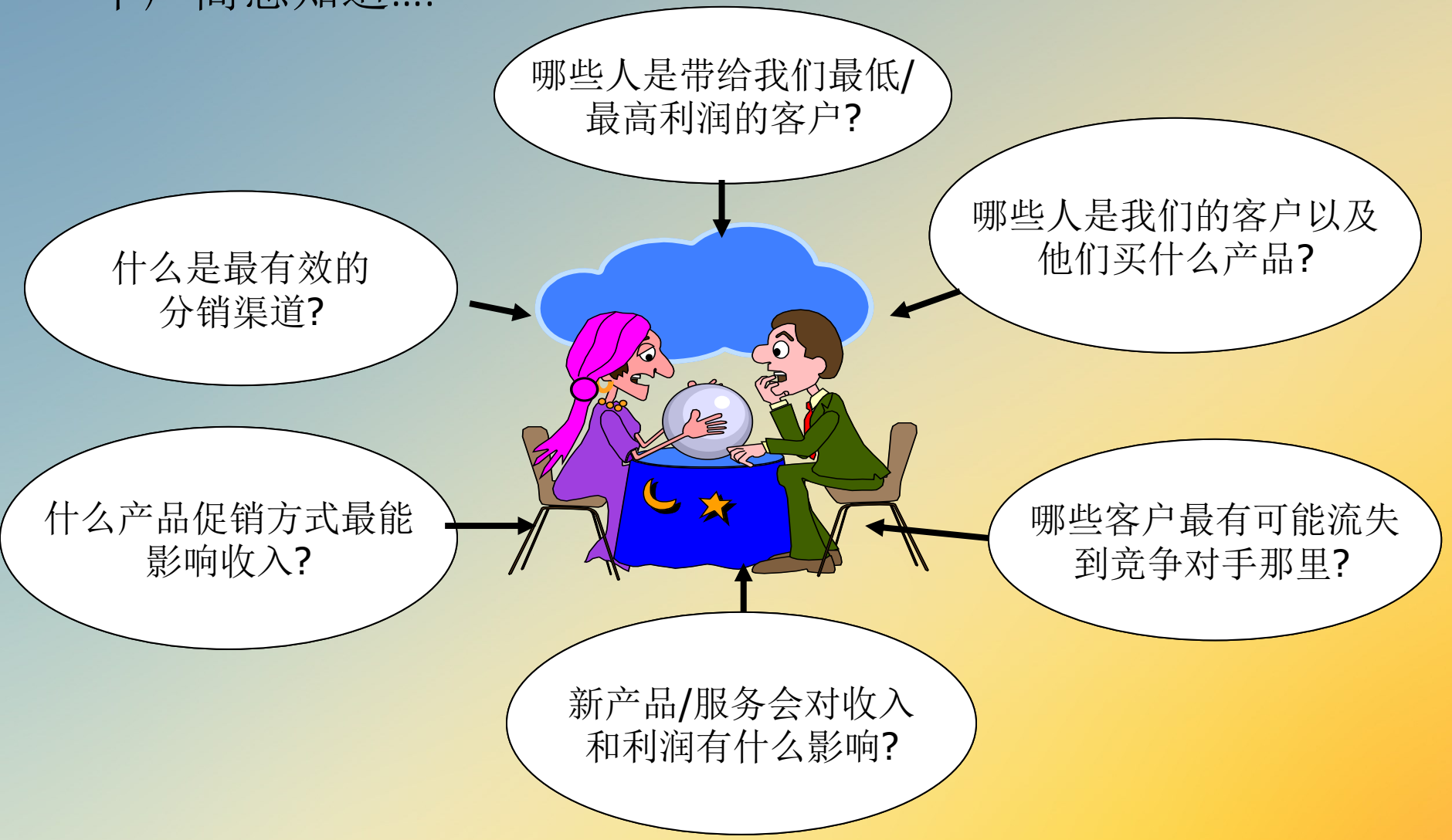
5.2 数据库系统与数据仓库

5.3 数据仓库的基本结构

5.4 数据仓库的相关概念

5.5 数据仓库与数据挖掘

一个厂商想知道....


哪些人是带给我们最低/
最高利润的客户?
哪些人是我们的客户以及
他们买什么产品?
什么是最有效的
分销渠道?
什么产品促销方式最能
影响收入?
哪些客户最有可能流失
到竞争对手那里?
新产品/服务会对收入
和利润有什么影响?

数据无处不在，然而 …

- 我找不到我所需要的数据
  - 数据分散在网络上的各个地方
  - 数据存在多个版本，其中有细小的差别
- 我不能获取我所需要的数据
  - 需要一个专家来获取数据
- 我无法理解所找到的数据
  - 可得到的数据，但对应的文档说明很糟糕
- 我无法使用所找到的数据
  - 结果不是期望的
  - 数据需要从一种形式转换到另外一种形式

演变过程

- 60年代: 批处理报表

  难于查找和分析信息

  缺乏灵活性，成本昂贵，对于每个新需求都要重新编程

- 70年代: 基于终端的EIS (主管信息系统)

  仍然缺乏灵活性，没有和桌面工具集成起来

- 80年代:桌面级数据访问和分析工具

  查询工具，电子表格，图形界面

  易于使用，但是只能访问操作型数据库

5.1.1数据仓库的产生

- **数据处理分为两类**
  - 事务处理
  - 分析处理
- **传统数据库较难满足分析处理的要求**
  - 历史数据需求量大
  - 不同系统的数据难以集成（蜘蛛网问题）
  - 对大量数据的访问性能不足
  - 事务处理和分析处理数据环境的分离

5.1.2 数据仓库的定义

- 数据仓库的定义
  - 20世纪80年代中期，“数据仓库”这个名词首次出现在号称“数据仓库之父”W.H.Inmon的《Building Data Warehouse》一书中，在该书中，W.H.Inmon把数据仓库定义为“一个面向主题的、集成的、稳定的、随时间变化的数据的集合，以用于支持管理决策过程。”（“A data warehouse is a *subject-oriented, integrated, non-volatile, time-variant* collection of data in support of management decisions.”）

5.1.2 数据仓库的定义

- William H.Inmon:数据仓库是一个面向主题的、集成的、非易失的且随时间变化的数据集合，用于支持管理人员的决策。

William H. Inmon: William H. Inmon是世界公认的“数据仓库之父”，是数据仓库及其相关技术网站**www.billinmon.com**的合作伙伴，是“企业信息工厂”的创造者之一。他一直致力于数据库和数据仓库技术方面的研究，在数据管理和数据仓库技术方面以及数据处理的管理方面撰写了**40**多本著作，发表过**600**多篇学术论文，并且经常应邀在技术和学术会议上演讲。

数据仓库之父

5.1.2 数据仓库的定义

通俗理解

数据仓库是一个将数据转换成信息、使其能及时供最终用户使用的过程

5.1.3 数据仓库的特征

- 数据仓库的特征
  - 数据仓库的数据是面向主题的
  - 数据仓库的数据是集成的
  - 数据仓库的数据是非易失的
  - 数据仓库的数据是随时间不断变化的

5.1.3 数据仓库的特征

- 主题（Subject）
  - 特定的数据分析领域与目标
- 面向主题
  - 为特定的数据分析领域提供数据支持
  - 数据仓库是面向分析、决策人员的主观要求的，不同的用户有不同的要求，同一个用户的要求也会随时间而经常变化
  - 数据仓库中的主题有时会因用户主观要求的变化而变化的。

5.1.3 数据仓库的特征

- **面向主题**
  - 特定主题的数据与传统数据库中的数据的差别
    - ✓ 传统数据库中的数据是原始的、基础的数据
    - ✓ 特定主题的数据则是需要对它们作必要的抽取、加工与总结而形成。

5.1.3 数据仓库的特征

- 例：一个面向事务处理的“商场”数据库系统，其数据模式如下
  - 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）

5.1.3 数据仓库的特征

- 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）

5.1.3 数据仓库的特征

- **面向主题的方式进行数据组织步骤**

1. 按照管理人员的分析要求来确定主题

   每个主题相关的数据又与有关的事务处理所需的数据不尽相同

2. 筛选包含有关该主题的相关信息，抛弃与分析该主题无关或不需要的数据

✓ 最终将原本分散在各个子系统中的有关信息集中在一个主题中，形成有关该主题的一个完整一致的描述

5.1.3 数据仓库的特征

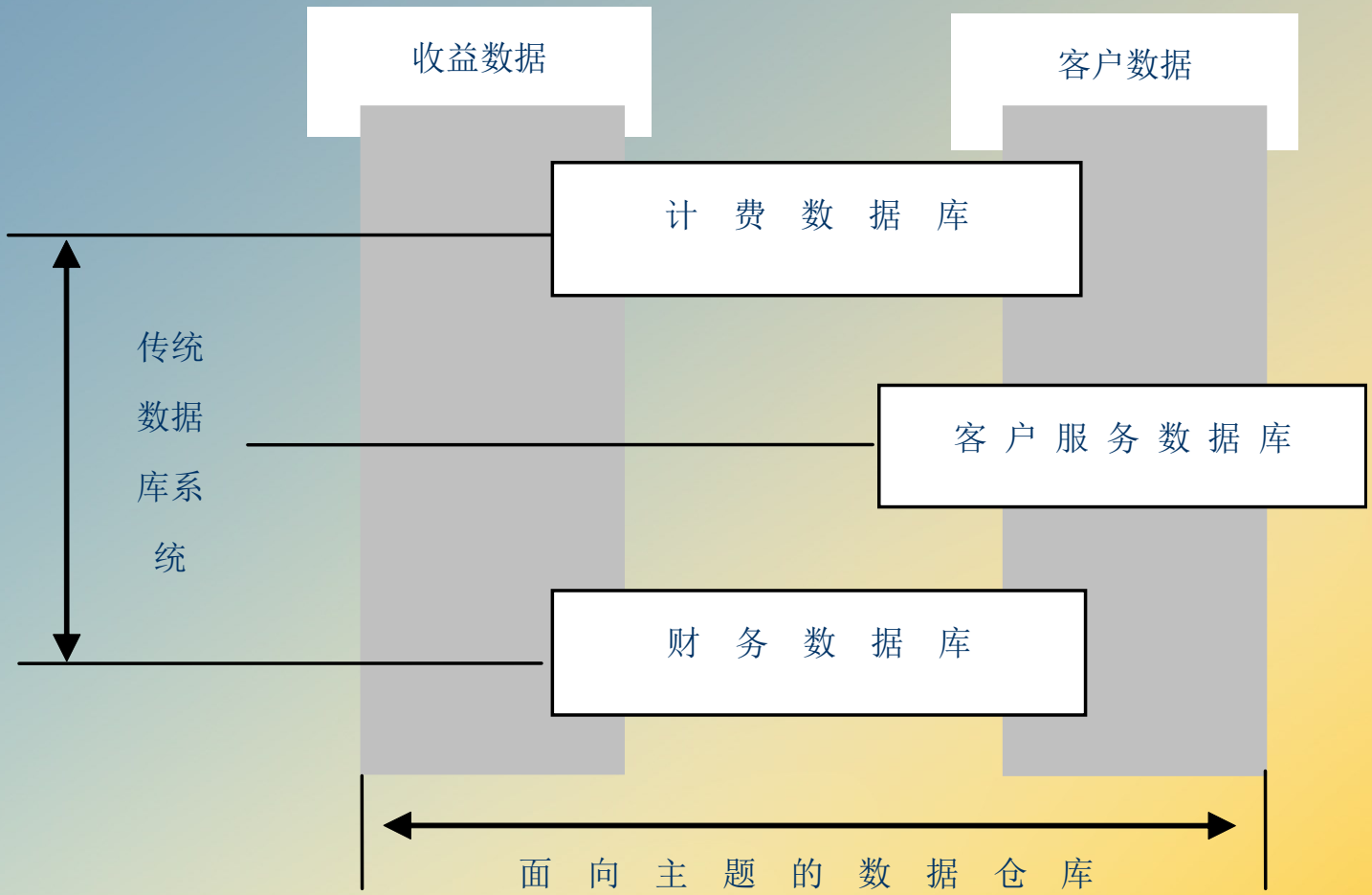
- **主题一：商品**
  - 商品固有信息：商品号，商品名，类别，颜色等
  - 商品采购信息：商品号，供应商号，供应价，供应日期，供应量等
  - 商品销售信息：商品号，顾客号，售价，销售日期，销售量等
  - 商品库存信息：商品号，库房号，库存量，日期等

5.1.3 数据仓库的特征

- 主题二：供应商
  - 供应商固有信息：供应商号，供应商名，地址，电话等
  - 供应商品信息：供应商号，商品号，供应价，供应日期，供应量等
- 主题三：顾客
  - 顾客固有信息：顾客号，顾客名，性别，年龄，文化程度，住址，电话等
  - 顾客购物信息：顾客号，商品号，售价，购买日期，购买量等

5.1.3 数据仓库的特征

- 一个电信企业的情况
  - 计费数据库：计费数据库记录了客户的消费情况
  - 财务数据库：财务数据库记录了客户的缴费情况
  - 客户服务数据库：客户的咨询和投诉情况



数据仓库面向主题的特性

5.1.3 数据仓库的特征

- 直接基于传统数据库系统进行“客户”和“收益”信息的分析
  - 需要访问多个数据库才能获得客户或收益各个侧面的信息，极大的影响系统处理的时间和效率
  - 数据之间的不一致性和不同步等问题将影响决策的可靠性
- 以“客户”和“收益”主题组织的数据仓库
  - 将某个主题的全部相关数据集中于一个地方
  - 决策者可以非常方便地在数据仓库中的一个位置检索包含某个主题的所有数据。

5.1.3 数据仓库的特征

- **面向集合（数据的集成）**
  - 数据仓库中的数据是为分析服务的

    数据仓库中的数据必须从多个数据源中获取
    - ✓ 多种类型数据库
    - ✓ 文件系统
    - ✓ Internet网上数据等

    通过数据集成而形成数据仓库中的数据
  - 这些数据源存在下述问题

    分散、数据不一致、外部数据和非结构化数据的问题

5.1.3 数据仓库的特征

- **面向集合**
  - 数据仓库中的数据是为分析服务的

不是原有数据的简单拷贝，而是经过了抽取、筛选、清理、综合等工作

对源数据的集成是数据仓库建设中最关键，也是最复杂的一步

5.1.3 数据仓库的特征

- 为什么不是简单地拷贝？
  - 数据仓库每一个主题所对应的源数据在源分散数据库中有许多重复或不一致之处
    - ✓ 必须将这些数据转换成全局统一的定义，消除不一致和错误之处
  - 源数据加载到数据仓库后，还要根据决策分析的需要对这些数据进行概括、聚集处理
  - 全面而正确的数据是有效地分析和决策的首要前提，相关数据收集得越完整，得到的结果就越可靠

5.1.3 数据仓库的特征

- 面向集合
  - 集成的方法：

    统一：消除不一致的现象

    综合：对原有数据进行综合和计算
  - 需要考虑的问题：

    数据格式

    计量单位

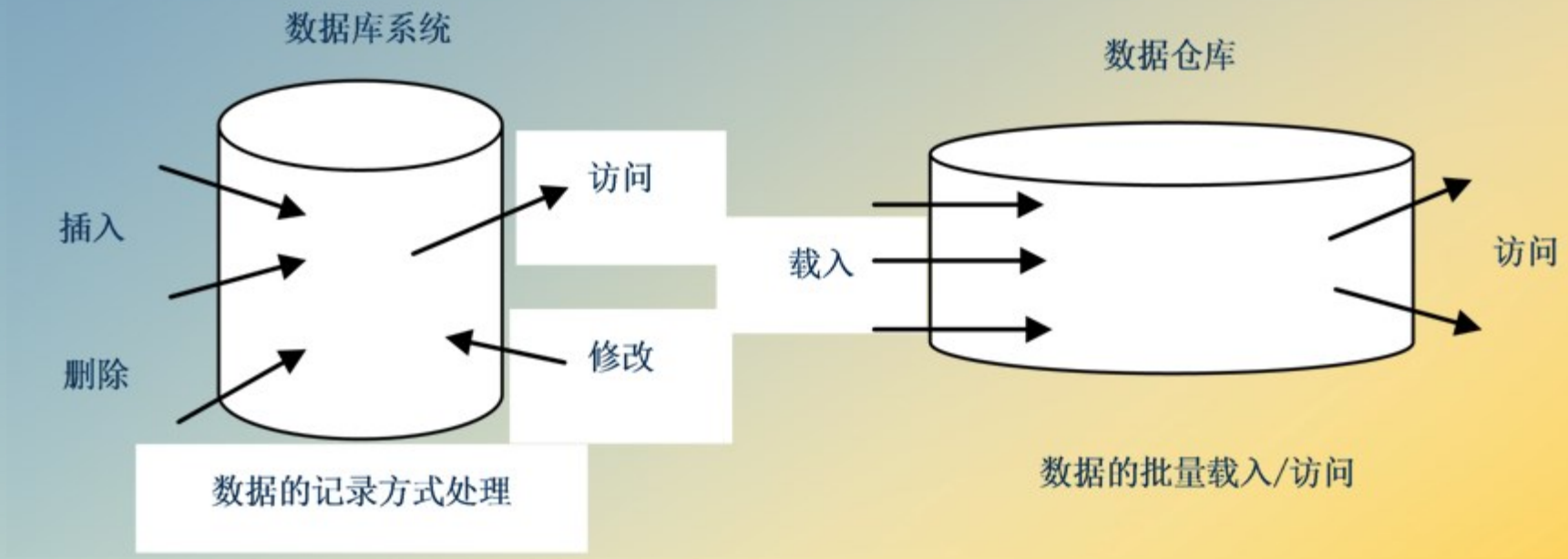
    数据代码含义混乱

    数据名称混乱

5.1.3 数据仓库的特征

- **非易失的**
  - 数据仓库中的数据是经过抽取而形成的分析型数据
    - ✓ 不具有原始性，主要供企业决策分析之用

      主要是‘查询’，一般不执行‘更新’
    - ✓ 一个稳定的数据环境也有利于数据分析操作和决策的制订
  - 不等于数据仓库中的数据不需要‘更新’操作。
    - ✓ 在需要进行新的分析决策时，可能需要进行新的数据抽取和‘更新’操作
    - ✓ 数据仓库中的一些过时的数据，也可以通过‘删除’操作丢弃掉。


数据库系统
数据仓库
插入
删除
访问
修改
载入
访问
数据的记录方式处理
数据的批量载入/访问

5.1.3 数据仓库的特征

- 随时间不断变化
  - 数据仓库中的数据必须以一定时间段为单位进行统一更新

    不断增加新的数据内容

    不断删去旧的数据内容

    更新与时间有关的综合数据

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的主要区别**

1、主要任务不同

2、数据内容不同

3、数据目标不同

4、数据特性不同

5、数据结构不同

6、支持的查询不同

7、数据组织模式不同

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的主要任务不同**
  - 传统数据库系统的主要任务是执行联机事务，即OLTP系统

    购买、库存、制造、银行、工资、注册、记帐

  - 数据仓库系统在数据分析和决策支持方面提供服务，即联机分析处理OLAP系统

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的数据内容不同**
  - 数据库系统管理当前数据。

    这种数据太琐碎，难以用于决策

  - 数据仓库系统管理大量历史的、存档的、归纳的、计算的数据，提供汇总和聚集机制，并在不同的粒度级别上存储和管理信息

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的数据目标不同**
  - 数据库系统是面向业务操作

    办事员、客户和信息技术专业人员的事务和查询处理
  - 数据仓库是面向主题的

    用于知识工人（包括经理、主管和分析人员）的决策分析

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的数据特性不同**
  - 数据库系统存储的是当前数据，数据是动态变化的，按字段进行更新操作
  - 数据仓库中数据是批量载入的、静态的，系统定期执行提取过程为数据仓库增加数据

    这些数据一旦加入，不会频繁地进行个别修改

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的数据结构不同**
  - 数据库系统采用面向应用的数据库设计，以高度结构化和复杂的形式组织数据，以适应事务操作计算的需求
  - 数据仓库通常采用面向主题的数据组织模式，以适应分析决策，数据结构简单

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统支持的查询不同**
  - OLTP系统提供了大量的原始数据，是为了应对快速回答、简单查询

    不是为了存储分析趋势的历史数据而创建的

    原始数据不易被分析
  - 数据仓库需要回答更复杂的查询
    - ✓ 不仅仅是一些像“这件商品多少钱？”之类的简单的数据查询
    - ✓ 数据仓库需要回答的查询类型可以是简单的查询，也可以是高度复杂的，且还与终端用户使用的查询工具相关
    - ✓ “每个分支机构本月的房产销售月收入是多少，并与刚过去的12个月相比较”
    - ✓ “如果对于10万英镑以上的房产，法定价格上升3.5%而政府税收下降1.5%，对英国不同区域的销售会产生什么影响？”

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的数据组织模式不同**
  - 数据库模式
    - ✓ 有关商品的信息分散在各个子系统之中
  - 数据仓库
    - ✓ 强调的就是要形成关于主题一致的信息集合
    - ✓ 丢弃了原来有的但不必要的、不适于分析的信息
    - ✓ 不同主题之间有重叠内容

5.2 数据库系统与数据仓库

- **数据仓库与传统数据库系统的数据组织模式不同**

传统数据库系统：

- 采购子系统
  - 定单（定单号，供应商号，总金额，日期）
  - 定单细则（定单号，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，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）
- 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，定单号，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，库房管理员，地点，库存商品描述）
- 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，部门主管，电话）

面向“商品”、“顾客”、“供应商”主题的数据仓库系统：

- 商品
  - 商品固有信息：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，供应量，供应日期等；
  - 商品销售信息：商品号，顾客号，售价，销售量，销售日期等；
  - 商品库存信息：商品号，库房号，库存量，日期等。
- 供应商
  - 供应商固有信息：供应商号，供应商名，地址，电话等；
  - 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
  - 顾客固有信息：顾客号，顾客名，性别，年龄，文化程度，地址，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。

5.2 数据库系统与数据仓库

数据库	数据仓库
面向应用	面向主题
数据是细节的	数据是综合的或提炼的
保存当前数据	保存过去和现在的数据
数据是频繁更新的	数据一般较少更新
一个操作存取少数数据	一个操作读取大量数据

5.2 数据库系统与数据仓库

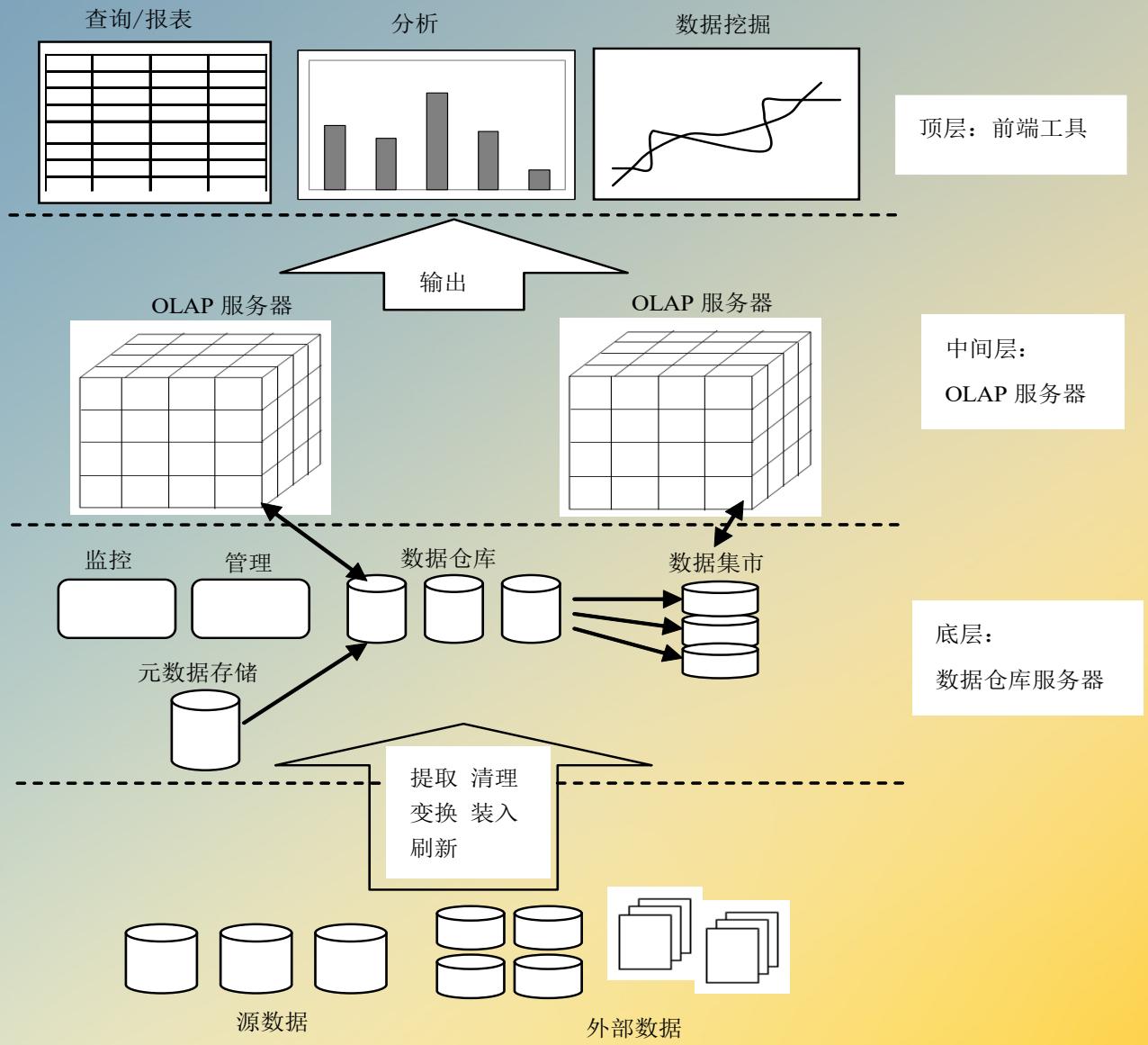
数据库	数据仓库
操作比较频繁	操作相对不频繁
查询的是原始数据	查询的是经过加工的数据
事务处理需要的是当前数据	决策分析需要现在、过去的数据
很少有复杂的计算	很多复杂的计算
支持事务处理	支持决策分析

5.3 数据仓库的基本结构

三层数据仓库结构

- 数据仓库服务器
- OLAP服务器
- 前端工具

三层数据仓库结构


查询/报表
分析
数据挖掘
顶层：前端工具
输出
OLAP 服务器
OLAP 服务器
中间层：
OLAP 服务器
监控
管理
数据仓库
数据集市
元数据存储
底层：
数据仓库服务器
提取 清理
变换 装入
刷新
源数据
外部数据

1） 底层是数据仓库服务器

- 一般为数据库系统
- 使用后端工具和实用程序从操作数据库和外部信息源加载和刷新它的数据

  ETL工具：具有数据抽取、数据清洗、数据转换、数据加载和数据刷新等功能
- 还包含一个元数据存储

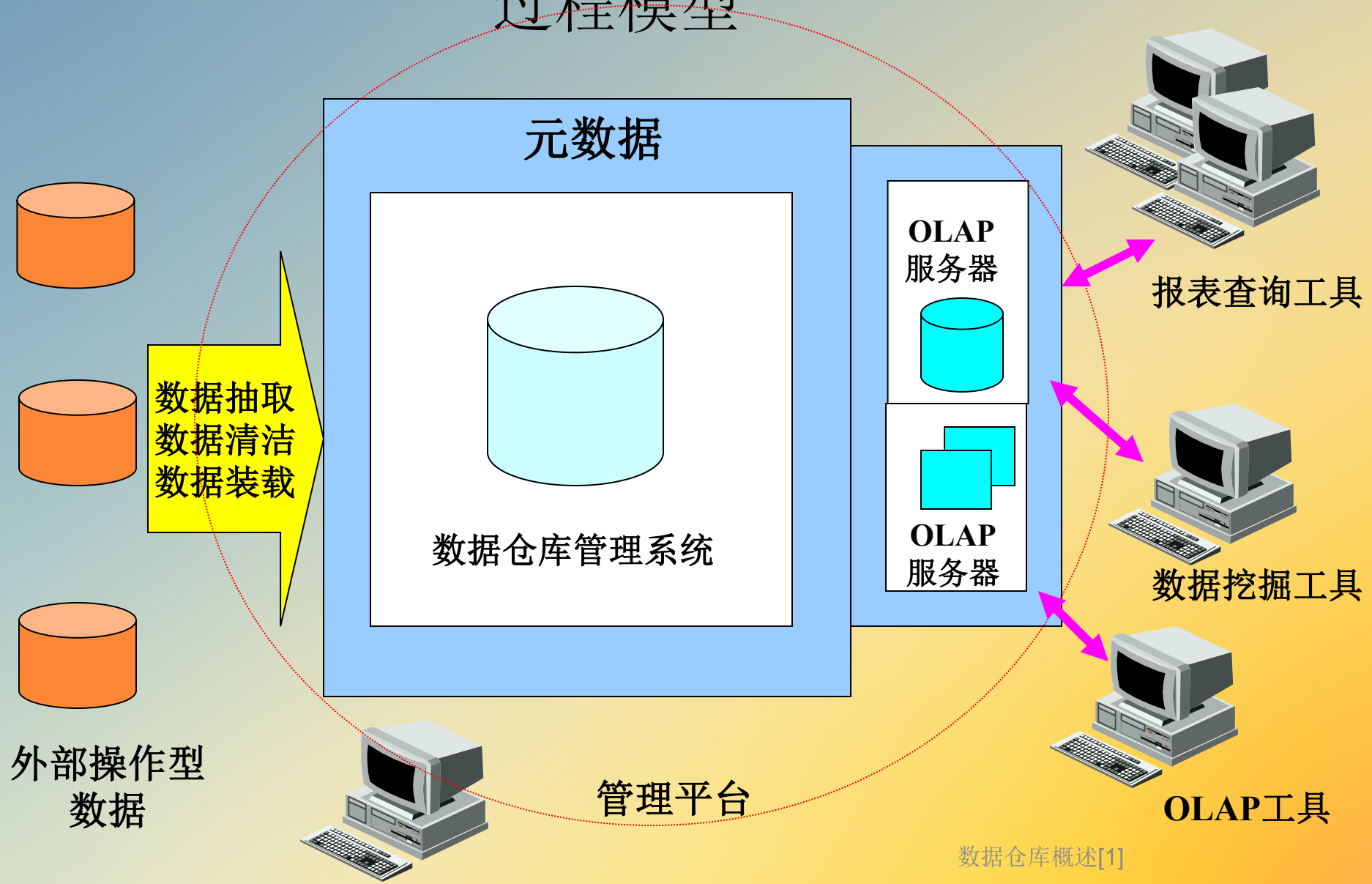
2） 中间层是OLAP服务器

- 存储数据，并实现多种数据操作的服务器

3） 顶层是客户

- 包括查询和报表工具、分析工具和/或数据挖掘工具（例如关联分析、分类分析、预测等）

过程模型


元数据
数据抽取
数据清洁
数据装载
数据仓库管理系统
OLAP
服务器
OLAP
服务器
报表查询工具
数据挖掘工具
OLAP工具
外部操作型
数据
管理平台

5.4 数据仓库的相关概念

- ETL
- 元数据（MetaData）
- 数据的粒度和维度
- 数据集市（Data Market）

5.4 数据仓库的相关概念

- ETL（Extract/Transformation/Load）—数据抽取、转换、加载工具
  - 数据提取（data extract）
  - 数据转换（data transform）
  - 数据清洗（data cleaning）
  - 数据加载（data loading）

5.4 数据仓库的相关概念

- 数据提取（Data Extract）
  - 数据仓库按照分析的主题来组织数据，只需提取出系统分析必需的数据
    - 例如，某超市以分析“客户的购买行为”为主题建立数据仓库，只需将与客户购买行为相关的数据提取出来，而超市服务员工的数据就没有必要放进数据仓库
  - 现有的数据仓库产品几乎都提供各种关系型数据接口，提供提取引擎，从关系型数据中提取数据

5.4 数据仓库的相关概念

- 数据转换（Data Transform）
  - 由于业务系统可能使用不同的数据库厂商的产品

    比如IBM DB2、Oracle、Informix、Sybase、 NCR Teradata、 SQL Server等
  - 各种数据库产品提供的数据类型可能不同
  - 需要将不同格式的数据转换成统一的数据格式

    如时间格式“年/月/日”，“月/日/年”、“日-月-年”的不一致问题等。

5.4 数据仓库的相关概念

- 数据清洗（Data Clean）
  - 对于决策支持系统来说，最重要的是决策的准确性
  - “清洗”就是将错误的、不一致的数据在进入数据仓库之前予以更正或删除，以免影响决策支持系统决策的正确性
  - 从多个业务系统中获取数据时，必须对数据进行必要的清洗，从而得到准确的数据

5.4 数据仓库的相关概念

- 数据加载（Data Load）
  - 数据加载部件负责将数据按照物理数据模型定义的表结构装入数据仓库
  - 包括清空数据域、填充空格、有效性检查等步骤

5.4 数据仓库的相关概念

- 元数据（MetaData）
  - **元数据**是描述数据的数据
    - ✓ 数据仓库中模型的定义、各层级间的映射关系、监控数据仓库的数据状态及 ETL 的任务运行状态
  - 元数据是数据仓库管理系统的重要组成部分，元数据管理是企业级数据仓库中的关键组件，贯穿了数据仓库的整个生命周期，使用元数据驱动数据仓库的开发，使数据仓库自动化，可视化
  - 目的是使数据仓库的设计、部署、操作和管理能达成协同和一致

5.4 数据仓库的相关概念

- 元数据主要包括
  a. 数据仓库结构的描述信息
  b. 操作元数据
  c. 汇总用的算法
  d. 由操作环境到数据仓库的映射信息
  e. 关于系统性能的数据信息
  f. 商务元数据

5.4 数据仓库的相关概念

- **数据仓库结构的描述信息**
  - 数据的维、层次结构、数据的定义
- **操作元数据**
  - 包括数据血统信息（来自何处以及如何转换的）
  - 数据流通信息
  - 监视信息（仓库使用统计、错误报告、审计跟踪）。

5.4 数据仓库的相关概念

- **汇总用的算法**
  - 包括度量与维定义算法，数据主题、聚集、汇总、预定义查询与报告的算法
- **由操作环境到数据仓库的映射信息**
  - 源数据库和它们的内容，ETL程序描述，数据分割、提取、清理和转换的规则，数据刷新和裁减的规则以及数据安全信息（用户授权和存取控制）

5.4 数据仓库的相关概念

- 关于系统性能的数据信息
  - 刷新、更新和复制周期的定时和调度的规则
  - 数据的索引
- 商务元数据
  - 包括商务术语和定义，数据拥有者信息和收费策略

5.4 数据仓库的相关概念

- 如果没有元数据
  - 数据仓库就像一个没有标签和文件夹的文件柜
  - 可能装满了很多对你的用户、开发者及管理者很有用的信息，却没有任何简便的方法知道这些信息在哪里，这样数据仓库的价值就很有限
- 在数据仓库中，元数据处于一个关键的位置，使不同的过程能够相互通信，是数据仓库的中枢。

实体名称：customer 别名：Account,Client

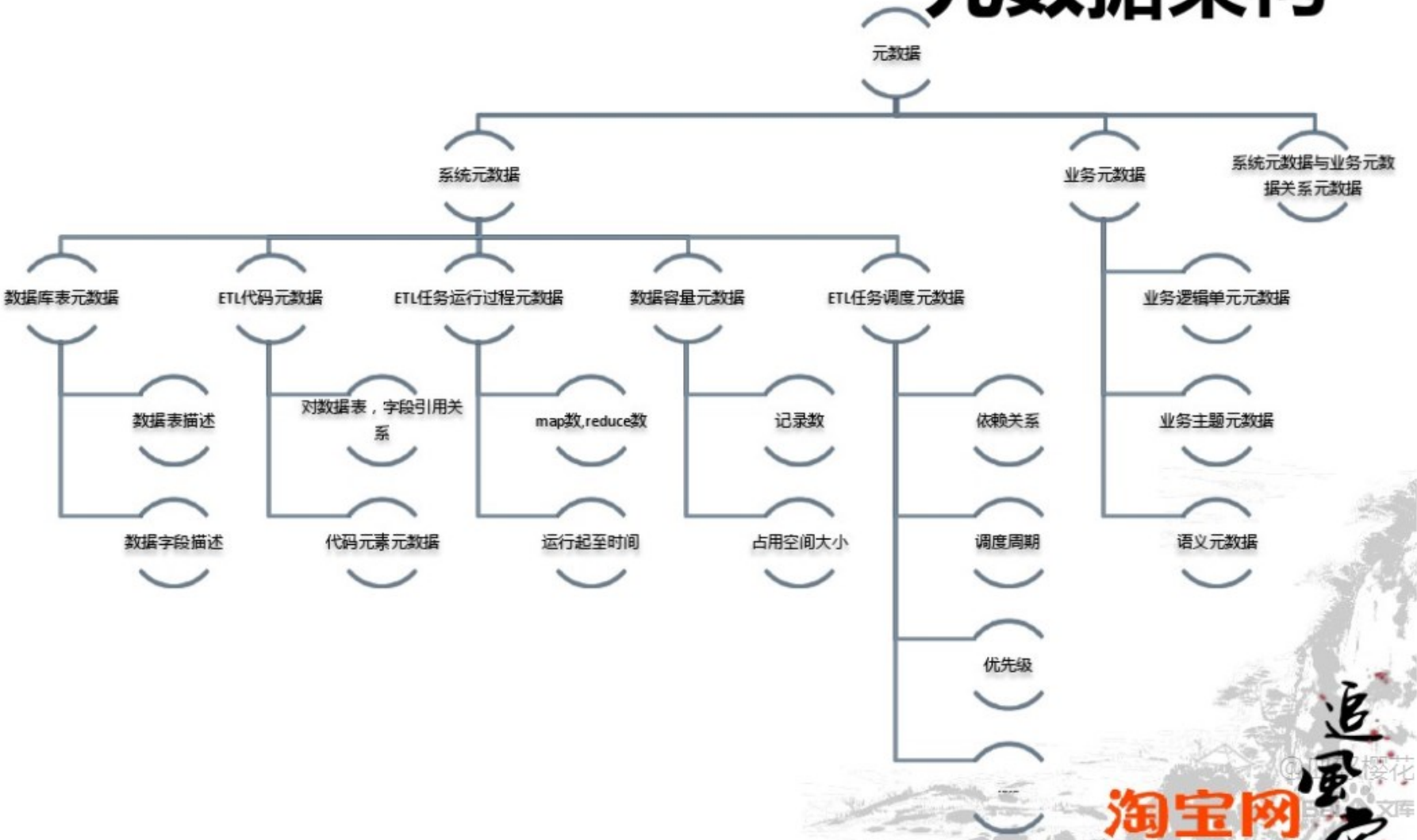
定义：从公司购买产品或服务的一个人或者一个机构。

备注：客户实体包含了常规的、当前以及过去的客户；

源系统：已经完成的产品订单，维护合同，在线销售

建立日期：	1999年1月15日
最后更新日期：	2001年1月21日
更新周期：	每周
最后的完全刷新日期：	2000年12月29日
完全刷新周期：	每6个月
数据质量回顾：	2001年1月25日
最后的副本：	2001年1月10日
计划归档：	每6个月
负责人：	jane brown

元数据架构


元数据
系统元数据
业务元数据
系统元数据与业务元数据关系元数据
数据库表元数据
ETL代码元数据
ETL任务运行过程元数据
数据容量元数据
ETL任务调度元数据
业务逻辑单元元数据
数据表描述
对数据表，字段引用关系
map数,reduce数
记录数
依赖关系
业务主题元数据
数据字段描述
代码元素元数据
运行起至时间
占用空间大小
调度周期
语义元数据
优先级
....

元数据在淘宝中的应用

5.4 数据仓库的相关概念

- 数据集市（Data Marts）
  - 一种更小、更集中的数据仓库，为公司提供分析商业数据的一条廉价途径
  - 通常是具有特定应用的数据仓库，主要针对某个应用或者具体部门级的应用
  - 数据集市的数据是从企业范围的数据库或者是更加专业的数据仓库中抽取出来的
  - 数据集市就是企业级数据仓库的一个子集

5.4 数据仓库的相关概念

- 建立数据集市的原因
  - 数据仓库是一种反映主题的全局性数据组织
  - 全局性数据仓库往往太大，在实际应用中将它们按部门或个人分别建立反映各个子主题的局部性数据组织，它们即是数据集市
  - 也称它为部门数据仓库。

例：在有关商品销售的数据仓库中可以建立多个不同主题的数据集市

商品采购数据集市

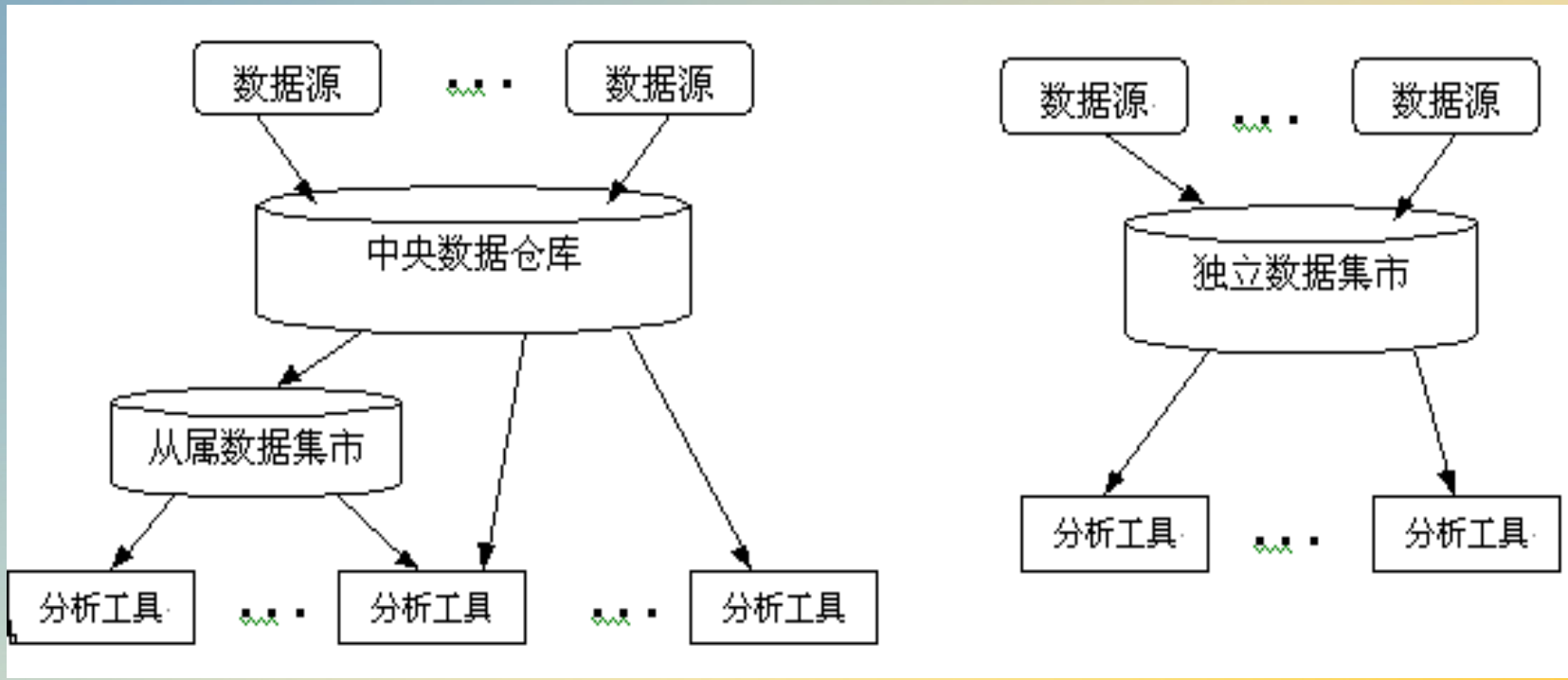
库房使用数据集市

商品销售数据集市

5.4 数据仓库的相关概念

- **数据集市的类型**
  - 独立数据集市(Independent Data Mart)
  - 从属数据集市(Dependent Data Mart)
- **按照数据获取来源：**
  - 独立型：直接从操作型环境获取数据。
  - 从属型：从企业级数据仓库获取数据。

5.4 数据仓库的相关概念



5.4 数据仓库的相关概念

- 数据仓库与数据集市的对比

	数据仓库	数据集市
数据来源	遗留系统、OLTP 系统、外部数据	数据仓库
范围	企业级	部门级或工作组级
主题	企业主题	部门或特殊的分析主题
数据粒度	最细的粒度	较粗的粒度
数据结构	规范化结构（第 3 范式）	星型模式、雪片模式、或两者混合
历史数据	大量的历史数据	适度的历史数据
优化	处理海量数据 数据探索	便于访问和分析 快速查询
索引	高度索引	高度索引

5.4 数据仓库的相关概念

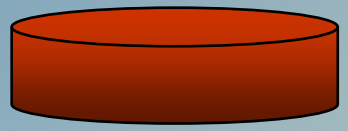
- **维度是人们观察数据的特定角度**
  - 一个企业在考虑产品的销售情况时，通常从时间、地区和产品的不同角度来深入观察产品的销售情况。
  - 时间、地区和产品就是维
  - 销售系统中的数据可分为时间维、产品维和地理位置维等；

5.4 数据仓库的相关概念

- 数据粒度
  - 对数据仓库中的数据综合程度高低的一个度量
  - 例如：每月交易的综合数据处于高粒度级，而每月所有交易的细节汇总则是低粒度级。
  - 粒度会深刻地影响存放在数据仓库中的数据量的大小以及数据仓库所能够回答的查询类型。

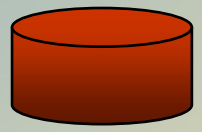
粒度

低粒度



一个顾客一个月中每次通话的细节

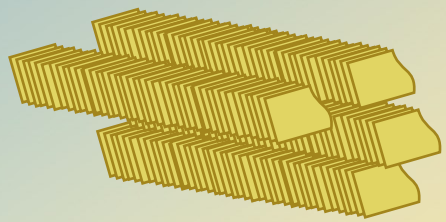
高粒度



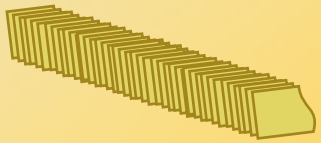
一个顾客一个月中通话的综合

Cass Squire上星期给他在波士顿的女友打过几次每次多长时间电话？

能回答，尽管需要一定数量的检索

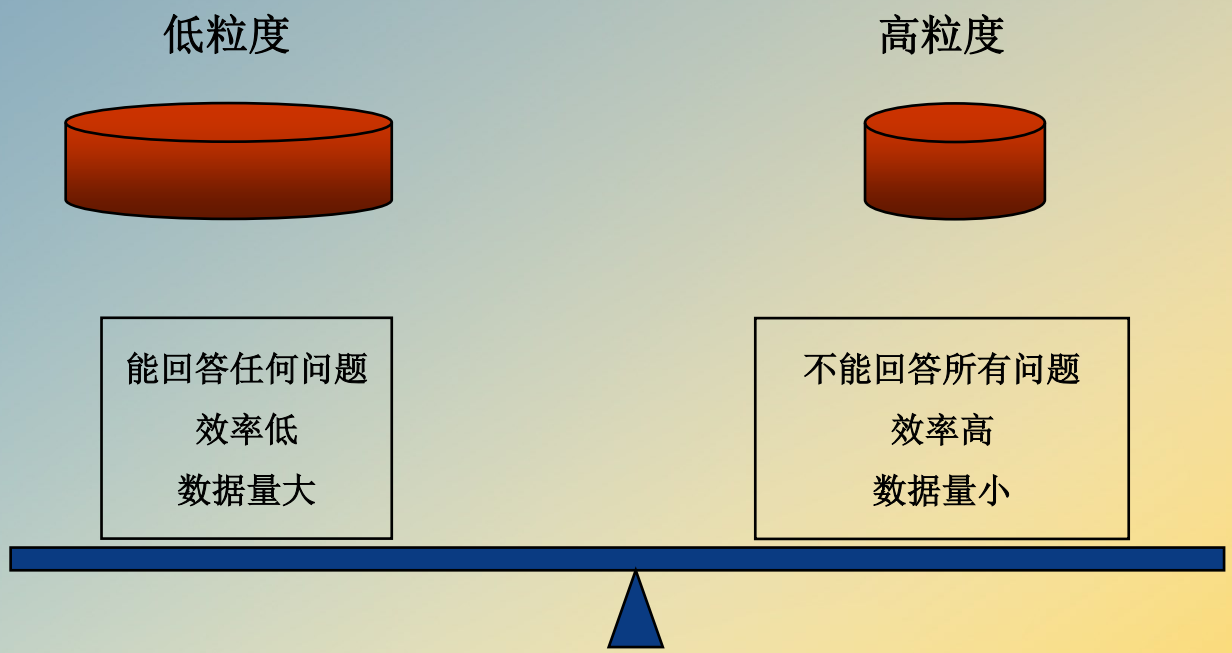


根本不能回答，细节已经丢失



由此可见，粒度级别对于能回答什么问题和问答问题所需资源多少有深刻的影响。

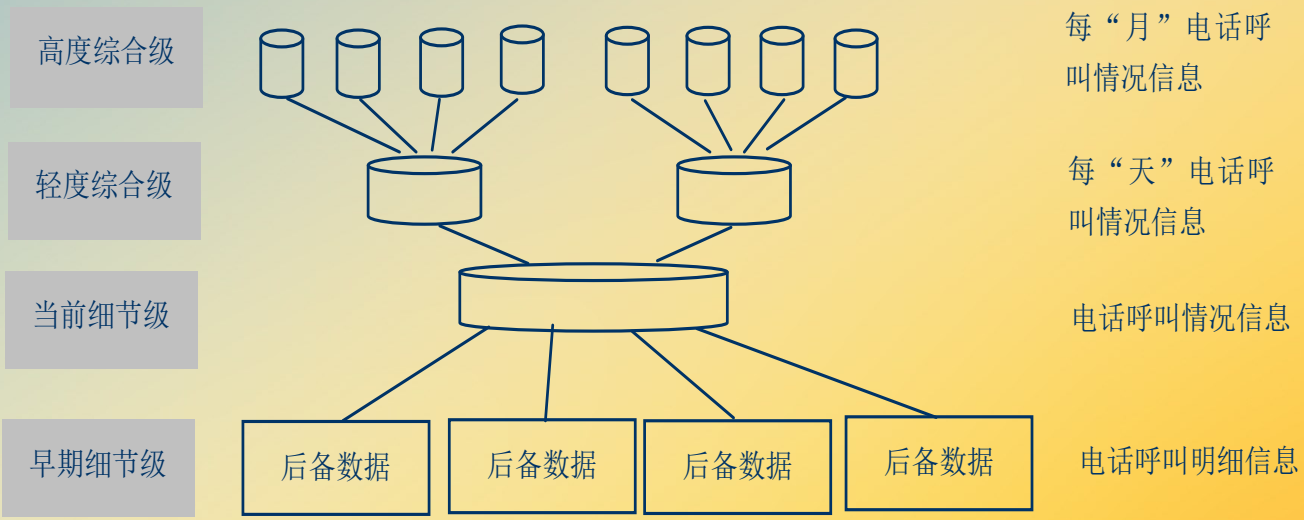
粒度



粒度的权衡是固有的，所以大多数企业的最佳解决方法是采用多重粒度的形式

5.4 数据仓库的相关概念

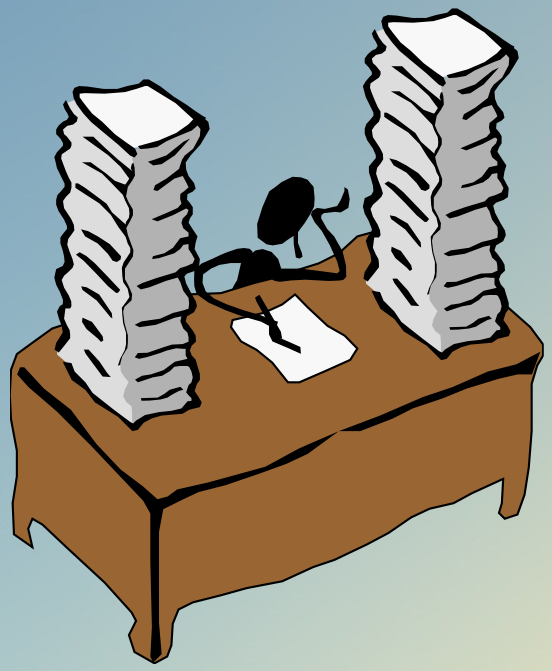
- 在数据仓库中，数据分成4个级别：
  - 高度综合级
  - 轻度综合级
  - 当前细节级
  - 早期细节级



5.4 数据仓库的相关概念

- **数据仓库的数据组织**
  - 数据仓库中存储着不同粒度的数据，粒度越大，表示细节程度越低，综合程度越高。
  - 进行OLAP分析时，常常需要不同层次的数据粒度
    - ✓ 通过预运算将数据综合成每个用户每“天”的通话次数，还可以进一步聚合成每个用户每“月”的通话次数。
  - 源数据（早期细节级数据）经过综合后，首先进入当前细节级，然后根据应用的需求，通过预运算将数据聚合成轻度综合和高度综合级
  - 轻度和高度是相对的概念，而没有绝对的界限，并且在数据仓库中数据的综合程度常常有很多的级别。

5.5 数据仓库与数据挖掘



- 数据仓库是企业的数据存储

- 数据挖掘是挖掘出企业数据中的知识

5.5 数据仓库与数据挖掘

- 如下领域的发展，使得数据挖掘的运用成为可能
  - 数据仓库自身的发展
  - 更好和更多的数据 (如, 操作型数据, 行为数据, 以及人口统计学数据)
  - 易于部署的数据挖掘工具的出现
  - 新的数据挖掘技术的出现

5.5 数据仓库与数据挖掘

◆数据挖掘是什么

知识发现（KDD）：从数据中发现有用知识的整个过程。

数据挖掘（DM）：KDD过程中的一个特定步骤，它用专门算法从数据中抽取知识。

如在人类数据库中挖掘知识为：

（头发=黑色）∨（眼睛=黑色）→亚洲人

该知识覆盖了所有亚州人的记录。

知识？ 人们对客观世界的规律性认识，数据很多，知识很少。

5.5 数据仓库与数据挖掘

- 数据仓库与数据挖掘的关系
  - 数据仓库与数据挖掘都是决策支持
  - 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。

5.5 数据仓库与数据挖掘

- 数据挖掘用于数据仓库实现决策支持
  - 预测客户购买倾向；
  - 客户利润贡献度分析；
  - 分析欺诈行为；
  - 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。

5.5 数据仓库与数据挖掘

- 数据挖掘从数据仓库中挖掘更深层次的信息
  - 哪些商品一起销售好？(关联分析)
  - 偏爱某类商品的客户特征是什么？（聚类分析）
  - 还有哪些客户具有上述特征？(类比分析)
  - 哪些商业事务可能具有欺诈性？（神经网络）
  - 高价值客户的共同点是什么？（分类分析）
  - 预测哪些用户可能流失？（分类分析）

5.5 数据仓库与数据挖掘

- **数据仓库为数据挖掘提出了新要求**
  - 数据挖掘需要可扩展性

    数据量逐渐递增
  - 数据挖掘方法需要能挖掘多维知识

    多维数据寻找它们的关联关系，例如商品与商店或时间等不同维之间的关联关系

总结

主要内容

5.1 数据仓库简介

5.1.1 数据仓库的产生

5.1.2 数据仓库的定义

5.1.3 数据仓库的特征

5.2 数据库系统与数据仓库

5.3 数据仓库的基本结构

5.4 数据仓库的相关概念

5.5 数据仓库与数据挖掘