上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

数据库和数据仓库

饶艳超 副教授

上海财经大学会计学院

raoyanchao@qq.com

学习目标

- 了解数据库及其管理系统的相关概念
- 熟悉数据仓库的定义和特征
- 熟悉数据仓库和数据库OLTP的主要区别

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

1

数据库及数据库管理系统

- **数据库术语清单**
  - 1. 数据库
  - 2. 表
  - 3. 列和数据类型
  - 4. 行
  - 5. 主键
  - 6. 查询和索引
    - ✓ 查询是人们用各种SQL指令构造出来的，SQL指令负责具体完成筛选和提取结果数据的工作。
    - ✓ 索引（index）是一种辅助性的数据表，它们只包含一种信息：原始数据记录的排序情况。

- **数据库管理系统**
  - 对数据库进行统一的管理和控制，以保证数据库的安全性和完整性。

    ✓用户通过DBMS访问数据库中的数据，

    ✓数据库管理员也通过DBMS进行数据库的维护工作。
  - 提供多种功能，可使多个应用程序和用户用不同的方法在同时或不同时刻去建立、修改和查询数据库。

- **数据库管理系统**
  - 对数据库进行统一的管理和控制，以保证数据库的安全性和完整性。

    ✓用户通过DBMS访问数据库中的数据，

    ✓数据库管理员也通过DBMS进行数据库的维护工作。
  - 提供多种功能，可使多个应用程序和用户用不同的方法在同时或不同时刻去建立、修改和查询数据库。

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

2

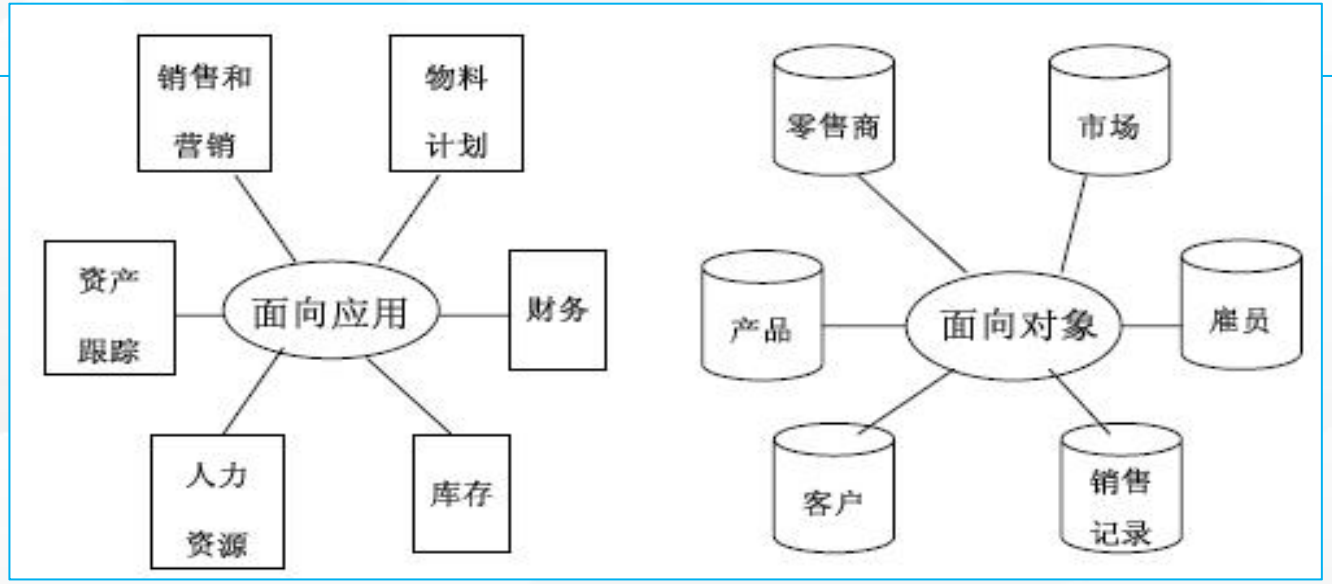
数据仓库

数据仓库

- 数据仓库的定义
  - 数据仓库领域的权威W.H.Inmon的定义：数据仓库是**支持管理决策过程**的-面向主题的、集成的、时变的、非易失的**数据集合**。
    - （1）**面向主题**(Subject Oriented)的数据集合
    - （2）**集成**(Integrated)的数据集合
    - （3）**时变**(Time Variant)的数据集合
    - （4）**非易失**(Non volatile)的数据集合

数据仓库

- 数据仓库的特性
  - 1.面向主题（Subject Oriented）：以用户需要的方式组织
    - ✓ 不同于面向功能（关注各种日常操作和事务处理）的各种应用程序，数据仓库面向决策支持（关注决策者的建模和分析）
    - ✓ 只需要考虑数据建模以及数据库的设计，无需顾及过程的设计
    - ✓ 数据之间相互联系



数据仓库

- 数据仓库的特性
  - 2. 数据集成（Integrated）：所有的名称和单位都进行了统一
    - ✓ 数据仓库中所有的数据都是整合的，是通过管理命名、度量属性、精确度和一般集合体的一致性表现出来的
    - ✓ 涉及应用程序的时候，对变量的命名是自由的，但是，一旦那些与应用程序相联系的数据库装载入数据仓库时，采用什么命名方式必须确定，需要使用统一的命名方式进行转换。
    - ✓ 数据整合的另一个结果是对**不同数据库中相似的数据**建立统一的单位，不仅仅要对**装入的数据**进行统一单位，而且还要对**最终数据**统一单位。
    - ✓ 数据之间相互联系

- 数据仓库的特性
  - 3. 时变性（Time Variant）——时间变量：不是当前的数据，而是时间序列数据，是反映历史变化情况的数据
    - ✓ 数据的时间变量有不同的表示方法
    - ✓ 数据的时间跨度比较长（5-10年），应用系统中的时间跨度是当前的或80-90天内
    - ✓ 另一个显示时间变量的地方是记录的主键，每个主键或显式或隐式的包含时间变量
    - ✓ 数据一旦被记录，将不可更改和变化。

- 数据仓库的特性
  - 4. 非易失（Non-Volatile）——即相对稳定、不可变性：只以只读的方式存储，不随时间变化
    - ✓ 在数据仓库中只有两种数据操作方法：数据装载和数据访问，以保证数据不可更改和更新
    - ✓ 应用设计，第三范式要求，无需存储所有可能的数据
    - ✓ 数据仓库，存储着很多操作数据中没有的计算结果和概括信息，非常有用

数据仓库

- **数据仓库的特性**
  - ✓5. 综合的：操作型数据映射为制定决策可以使用的格式
  - ✓6. 海量的：时间序列数据集一般数据量很大
  - ✓7. 元数据：关于数据存储的数据
  - ✓8. 数据源：数据来自内部和外部的未经过整合的操作型系统

数据仓库

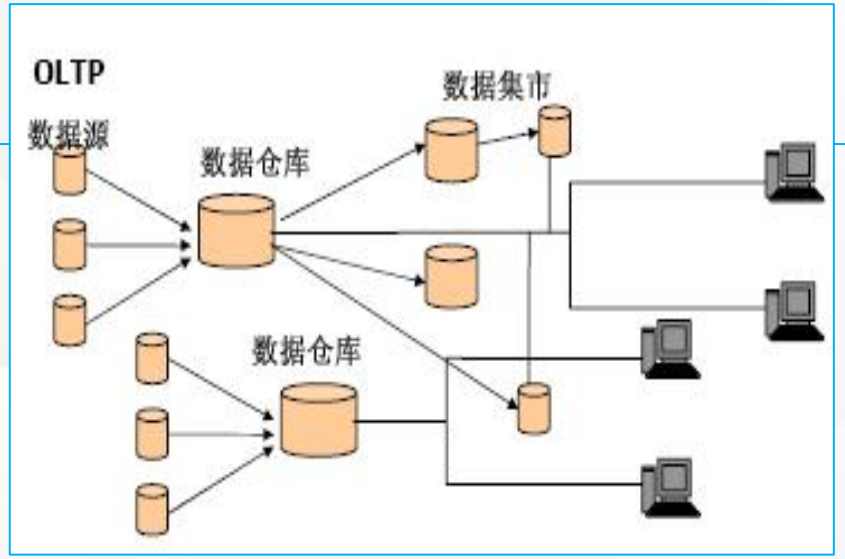
- **数据仓库的特性**

✓数据仓库的目的是**构建面向分析的**集成化数据环境，为企业提供决策支持。

✓数据仓库本身不“生产”任何数据，同时自身也不需要“消费”任何数据。

✓数据来源于外部，并且开放给外部应用。

✓数据仓库的基本架构主要描述数据流入流出的过程，分为三层：源数据、数据仓库、数据应用。



数据仓库

- **数据仓库中的数据包括：当前详细数据、历史数据、元数据。**

✓ **1、当前详细数据：**

- ✓ 数据仓库的核心，存放大量数据。
- ✓ 数据来自业务操作数据库，通过主题来组织，不是代表特定应用，而是代表整个企业。
- ✓ 在仓库中数据粒度最低，当数据精确化时，其中的每一个数据实体都是一个快照、一个时刻，表示一个瞬间。
- ✓ 一旦需要经常支持企业需求，数据随即进行更新。

数据仓库

- **数据仓库中的数据包括：当前详细数据、历史数据、元数据。**

✓ **2、历史数据：**

- ✓ 以前的有意义数据（一般两年以上），给企业带来延续的利益和价值。
- ✓ 包含巨大的数据量，可以用来预测和趋势分析。
- ✓ 包括：旧数据（原始或汇总形式）、描述旧数据特征的元数据。

✓ **3、元数据：**

- ✓ 关于数据的数据，也称为数据仓库的结构，是所有数据的集成体现。
- ✓ 数据仓库的最重要的部分，仓库开发者使用元数据来管理和控制仓库的建立和维护。

数据仓库

●**数据粒度：**用于定义数据仓库所存储信息的概要程度。

✓不同粒度表示为不同级别的汇总数据。

✓汇总数据是数据仓库的特点，所有的企业需要的数据分类（按部门、地区、功能等）及汇总程度都不同，数据仓库一般会提供不同程度的汇总数据。

✓轻量级汇总数据为整个企业组成部分服务。通过企业数据分类找到详细和汇总数据。但是它依旧比仓库中的详细数据少得多。

✓高度汇总数据是企业执行的主要依据，它来自根据企业组成部分的轻量级汇总数据或来自当前详细数据。这一层的数据容量比其他任何一个都少，代表一个折衷的积累，用来支持广泛的各式的需要和兴趣。通过高度汇总，执行者能够使用“钻取”到达逐步增加的详细层。

数据仓库

●数据仓库不需要存储所有原始数据，同时数据仓库需要存储部分细节数据。

✓为什么不需要存储所有原始数据？

✓成本效益分析原则：数据仓库面向分析处理，但是某些源数据对于分析而言没有价值或者其可能产生的价值远低于储存这些数据所需要的数据仓库的实现和性能上的成本。

✓例如：对销售而言，知道用户的省份、城市足够，至于用户究竟住哪里可能只是物流关心的事。

✓对物流而言，用户在博客中评论的内容只有与物流配送有关的才有意义，把所有评论文本都保存在数据仓库中可能得不偿失。

数据仓库

- 数据仓库不需要存储**所有原始数据**，同时数据仓库需要存储**部分细节数据**。
  - ✓**为什么要存储细节数据？**
    - ✓细节数据是必需的，数据仓库的分析需求会时刻变化，而有了细节数据就可以做到以不变应万变。
    - ✓如果我们只存储根据某些需求搭建起来的数据模型，那么显然对于频繁变动的需求会手足无措。

数据仓库

●数据仓库如何在维护细节数据的基础上对数据进行处理，使其能够真正应用于分析？

- ✓ 1. 数据的聚合
- ✓ 聚合数据指的是基于特定需求的简单聚合（基于多维数据的聚合体现在多维数据模型中）。
- ✓ 简单聚合可以是网站的总页面浏览量、访问数、独立访客等汇总数据，也可以是页面平均停留时间、站点平均访问时间等平均数据。
- ✓ 这些数据可以直接地展示于报表上。

●数据仓库如何在维护细节数据的基础上对数据进行处理，使其能够真正应用于分析？

✓ 2. 多维数据模型

✓ 多维数据模型提供多角度多层次的分析应用，比如基于时间维、地域维等构建的销售星形模型、雪花模型，可以实现在各时间维度和地域维度的交叉查询，以及基于时间维和地域维的细分。

✓ 数据仓库面向特定群体的数据集市都是基于多维数据模型构建的。

数据仓库

●数据仓库如何在维护细节数据的基础上对数据进行处理，使其能够真正应用于分析？

✓ 3. 业务模型

✓ 指的是基于某些数据分析和决策支持而建立起来的数据模型。如用户评价模型、关联推荐模型、RFM（Recency-Requency-Montary）分析模型、线性规划模型、库存模型等；

✓ 数据挖掘中前期数据的处理也可以借助业务模型完成。

数据仓库

- 数据仓库的数据应用

✓**报表展示：**报表几乎是每个数据仓库的必不可少的一类数据应用，将聚合数据和多维分析数据展示到报表，提供了最为简单和直观的数据。

✓**即时查询：**理论上数据仓库的所有数据（包括细节数据、聚合数据、多维数据和分析数据）都应该开放即时查询，即时查询提供了足够灵活的数据获取方式，用户可以根据自己的需要查询获取数据。

数据仓库

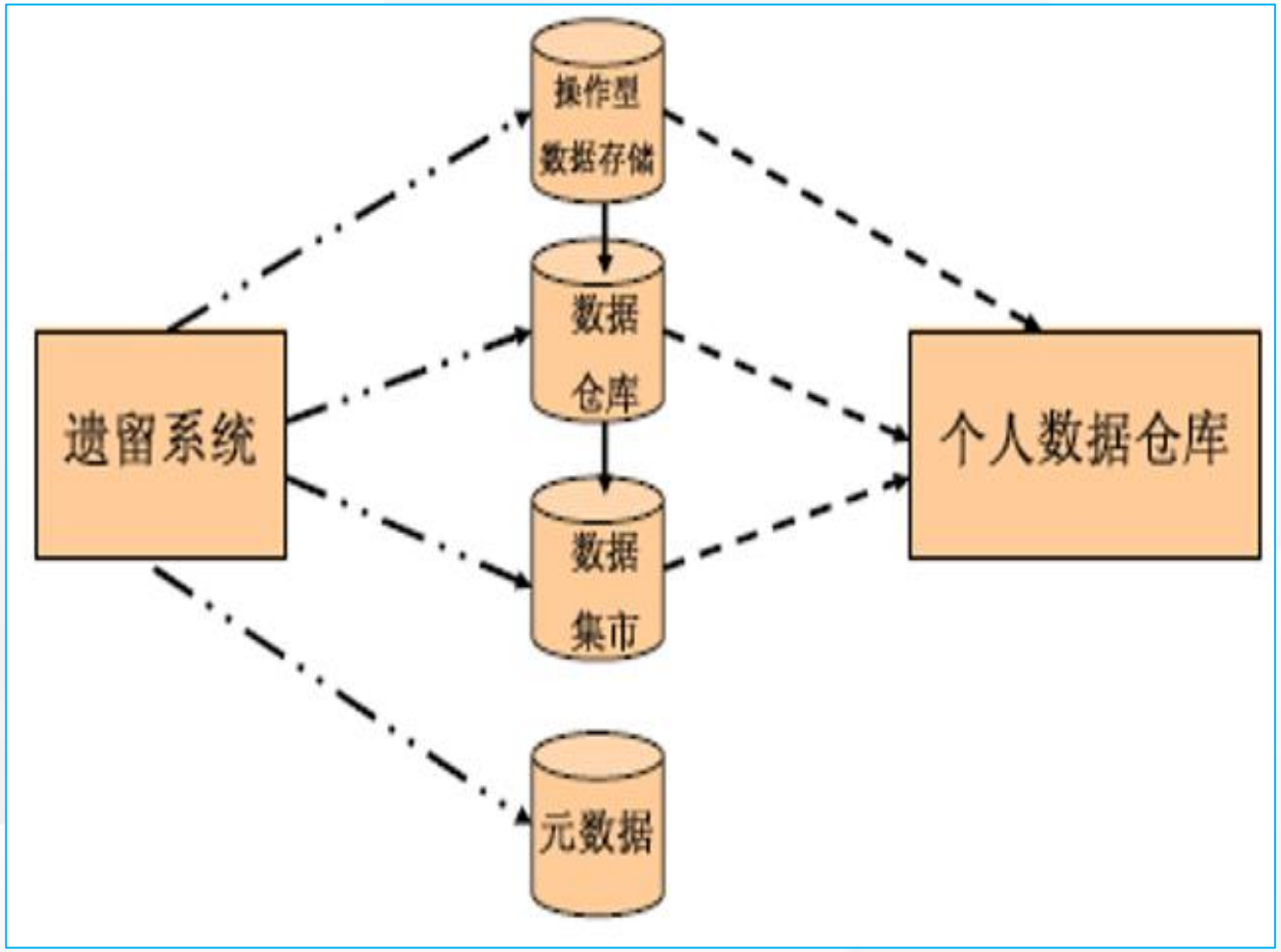
- **数据仓库的数据应用**

✓**数据分析：**大部分基于构建的业务模型展开，当然也可以使用聚合的数据进行趋势分析、比较分析、相关分析等，而多维数据模型提供了多维分析的数据基础；同时从细节数据中获取一些样本数据进行特定的分析也是较为常见的一种途径。

✓**数据挖掘：**数据挖掘用一些高级的算法可以让数据展现出各种令人惊讶的结果。数据挖掘可以基于数据仓库中已经构建起来的业务模型展开，但大多数时候数据挖掘会直接从细节数据上入手，而数据仓库为挖掘工具诸如SAS、SPSS等提供数据接口。

数据仓库

- 数据仓库的类型



数据仓库

- **操作型数据存储(ODS)**
  - ✓ 数据仓库环境中最基本的组成部分
  - ✓ 每天存储各种应用程序的数据
  - ✓ 为数据仓库提供必需的原始数据
  - ✓ 数据组织形式是面向对象的（顾客、产品、订单、政策等）、易变的、近期的
  - ✓ ODS通常来源于一个或多个遗留系统
    - ✓ 遗留系统在企业中广泛存在，主要指那些过时或存在问题的计算机系统
    - ✓ 为了能够用于分析，都必须进一步整合到数据仓库中

数据仓库

- 操作型数据存储(ODS)
  - ✓ 具备数据仓库的部分特征和OLTP系统的部分特征
  - ✓ 是“集成的、当前或接近当前的、不断变化的”数据；
  - ✓ 一般不保留数据变动轨迹，是数据仓库体系结构中的一个可选部分；
  - ✓ 主要是快速采集源数据；
  - ✓ 一般也会采用DW的一些技术；
  - ✓ 可以部分保留较少天数的历史数据，不能满足企业的中远期决策需求；
  - ✓ 没有稳定的数据层；

数据仓库

- 操作型数据存储(ODS)
  - ✓ ODS是面向主题和面向综合的；
  - ✓ ODS是易变的，ODS仅仅含有目前的、详细的数据，不含有累计的、历史性的数据。
  - ✓ 一般ODS用于报表数据源，同时为DW提供数据；
  - ✓ DW作决策支持，提供历史数据；

数据仓库

- 企业数据仓库
  - ✓ 通用数据仓库
  - ✓ 既含有大量详细的数据，也含有大量累赘的或聚集的数据
  - ✓ 数据具有不易改变性和面向历史性。

数据仓库

- 数据集市

✓一个小型的部门或工作组级别的数据仓库

✓也可叫做“小数据仓库”，是数据仓库的一种具体化。

✓有两种类型的数据集市——独立型和从属型。

- 独立型数据集市直接从操作型环境获取数据。
- 从属型数据集市从企业级数据仓库获取数据。
- 从长远的角度看，从属型数据集市在体系结构上比独立型数据集市更稳定。

数据仓库

- **数据集市**

✓主要面向部门级业务，并且只面向某个特定的主题

✓可以包含轻度累计、历史的部门数据，适合特定企业中某个部门的需要。

✓如果说数据仓库是建立在企业级的数据模型之上的话，那么数据集市就是企业级数据仓库的一个子集。

✓数据集市可以在一定程度上解决访问数据仓库的瓶颈。

数据仓库

- **数据集市**

✓适用于构建小型的、低成本的数据仓库。

✓几组数据集市可以组成一个企业数据仓库。

✓如果在整个企业的层次上构筑，可以提供低成本的数据存储并不断扩大发展成为整个的数据仓库环境。

✓常被视为开发数据仓库的一种方法，直接向一个独立的数据使用者提供数据更为容易。

✓不能够从企业的范围内进行规划，数据集市成为一个个信息孤岛。

数据仓库

- 元数据
  - ✓是数据的数据
  - ✓是关于数据仓库的信息，而不是数据仓库内存储的信息
  - ✓描述数据仓库中存储了什么样的数据、存储的位置，如何获得数据等方面的内容。
  - ✓元数据是数据仓库的核心，它用于存储数据模型和定义数据结构、转换规划、仓库结构、控制信息等。
  - ✓包括数据源元数据、ETL规则元数据、OD元数据、报表元数据、接口文件元数据、业务规则元数据等。

数据仓库

- 元数据
  - ✓数据仓库中存了什么表、属性和键？
  - ✓每一个数据集合的来源是什么？
  - ✓在数据装载入库时使用的什么转换逻辑？
  - ✓元数据如何随时间变化？
  - ✓数据的别名是什么以及数据之间的关系如何？
  - ✓技术和业务过程的关联是什么？
  - ✓数据重载的频率是多少？
  - ✓数据仓库中共有多少数据元素？……

- 元数据
  - ✓程序员所知的数据结构
  - ✓DSS分析员所知的数据结构
  - ✓数据仓库的源数据
  - ✓数据加入数据仓库时的转换
  - ✓数据模型
  - ✓数据模型和数据仓库的关系
  - ✓抽取数据的历史记录
  - ✓……

数据仓库

- 元数据

✓元数据是高层次的数据，提供对低层次数据的简明的抽象。

✓数据仓库中的数据不论是不是元数据都是不可更新的。

✓数据仓库中数据的每一次增加，元数据都会进行扩展

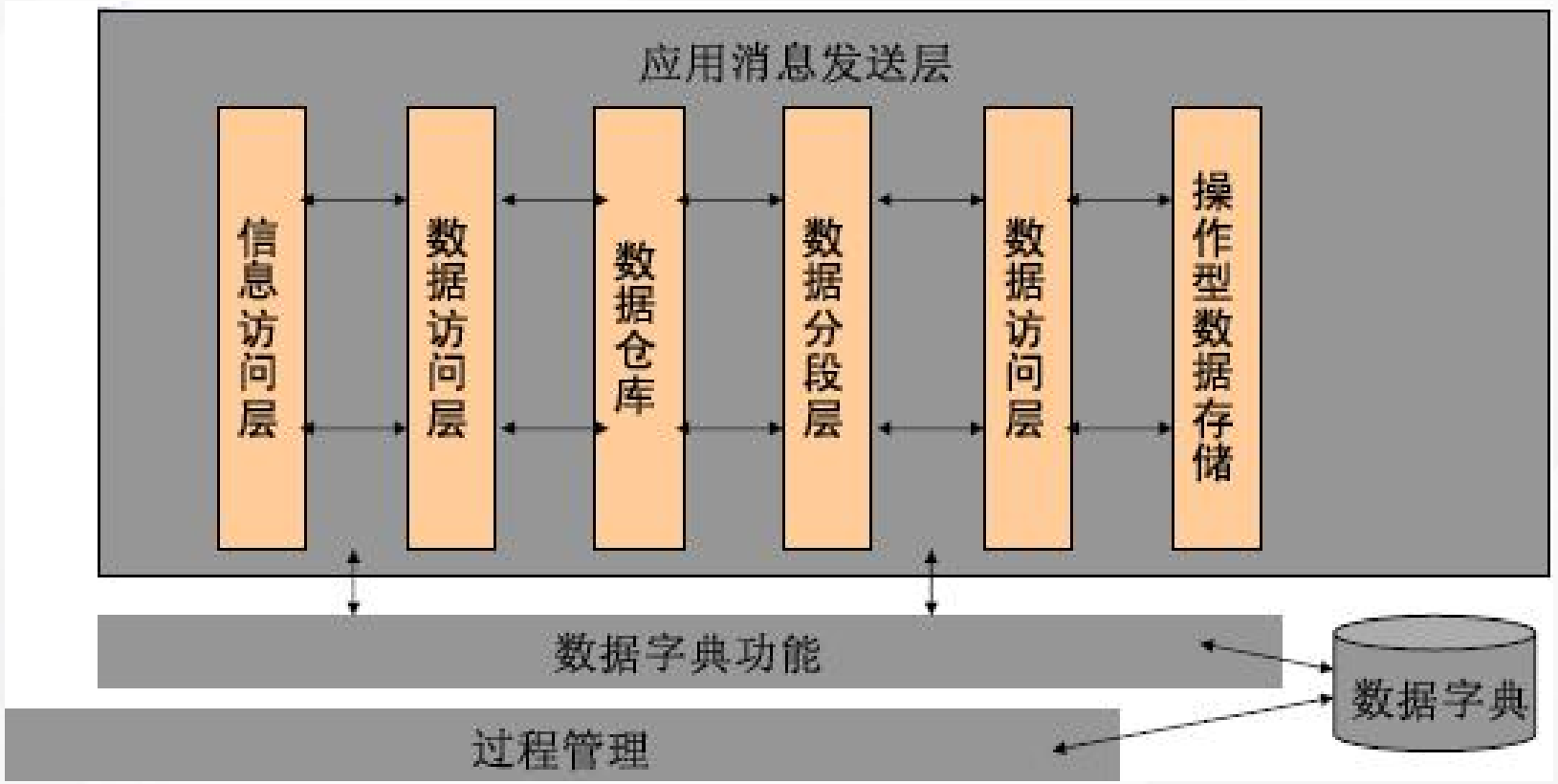
✓为了描述数据仓库中大量的元素，元数据必须组织为精确的、前后对照的方式。

数据仓库

- 数据仓库系统
  - （1）数据仓库：
    - ✓数据仓库的数据来源于多个数据源,包括企业内部数据、市场调查报告及各种文档之类的外部数据。
  - （2）仓库管理：
    - ✓在确定数据仓库信息需求后，首先进行数据建模，然后确定从数据源到数据仓库的数据抽取、清理和转换过程，最后划分维数及确定数据仓库的物理存储结构。
    - ✓仓库管理包括对数据的安全、归档、备份、维护、恢复等工作,这些工作需要利用数据库管理系统(DBMS)的功能。
  - （3）分析工具：
    - ✓完成决策问题所需的各种查询检索工具、多维数据的OLAP分析工具、数据开采DM工具等。

数据仓库

- 数据仓库的应用层级



数据仓库

- **数据仓库的应用层级**
  - 操作和外部数据库层：数据仓库的数据源，用户不必考虑访问数据库的操作型应用的执行过程。
  - 信息访问层：直接与最终用户打交道的一层，最终用户用来提取和分析数据仓库中数据的工具。
  - 数据访问层：连接操作型信息访问层与数据仓库本身的一个接口，包括数据仓库所涉及的不同的数据库，为数据仓库用户访问数据提供方便。
  - 数据分段运输层：包括选择、编辑、小结、合并以及从操作性和/或外部数据库中装载数据仓库和信息访问数据的所有过程。

数据仓库

- 数据仓库的应用层级
  - 应用消息发送层：
    - 用于在计算机网络中传递信息。
    - 不仅包括网络协议和请求路由的功能，还可以使得操作和信息的应用于数据的格式相隔离。
    - 可视作数据仓库底层的传输系统。
  - 物理数据仓库层：数据实际存储的地方，包括虚拟的和本地的数据。
  - 元数据层：为实现通用的数据访问服务。
  - 过程管理层：主要着重于调度数据仓库的建立以及元数据的维护所必需的各种任务。

数据仓库

- 数据仓库的建设
  - 面向主题

✓从公司业务出发，是分析的宏观领域，比如供应商主题、商品主题、客户主题和仓库主题

  - 为多维数据分析服务

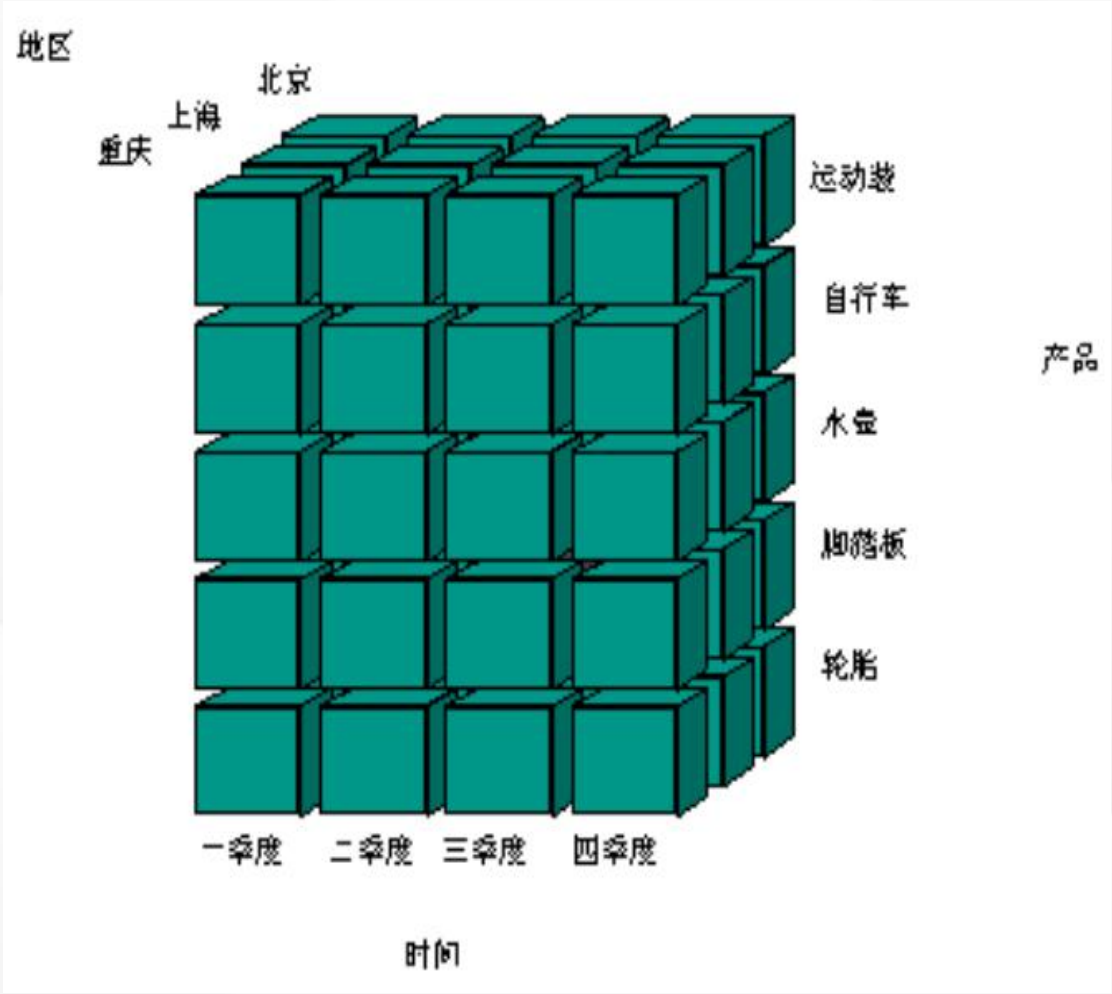
✓数据报表；数据立方体，上卷、下钻、切片、旋转等分析功能。

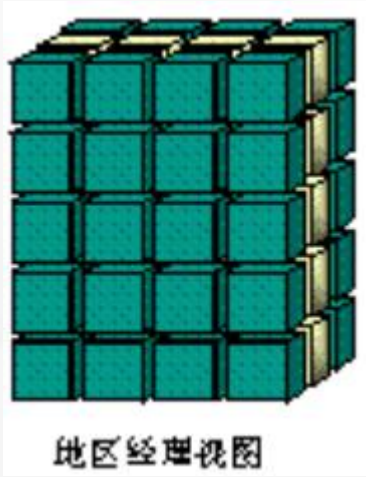
  - 反范式数据模型

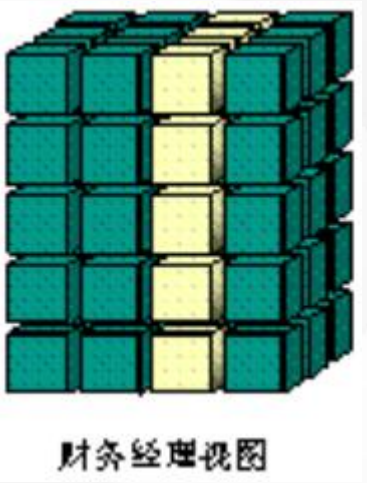
✓以事实表和维度表组成的星型数据模型

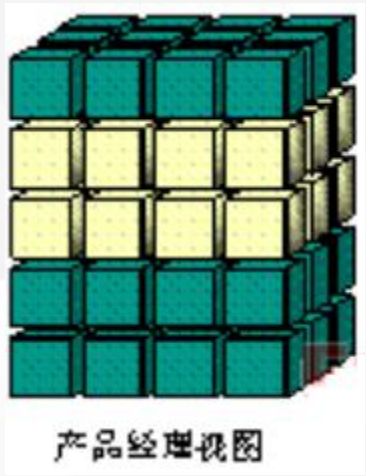
数据仓库

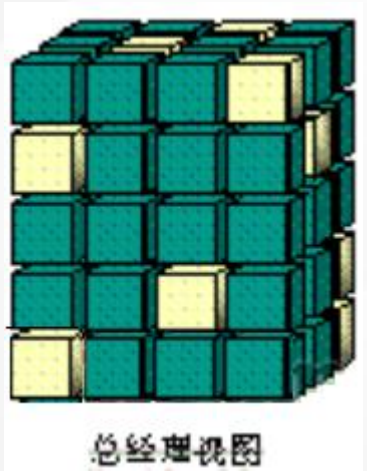
- 星型数据模型




地区经理视图


财务经理视图


产品经理视图


总经理视图

数据仓库

- **数据仓库与数据库、OLTP的区别**
- 以银行业务为例：
  - ✓ 数据库是在线事务系统的数据管理平台，如AMT系统，客户在银行做的每笔交易都会写入数据库，被记录下来。
  - ✓ 数据仓库是分析系统的数据平台，从事务系统获取数据，并做汇总、加工，为决策者提供决策的依据。如，某银行某分行ATM的交易数据，用于分析分行的ATM使用效率。

数据仓库

- **数据仓库与数据库OLTP的区别**
- 以Email为例
  - Email 程序是一个OLTP（在线事务处理系统），需要一个不是很复杂的数据库支持。
    - ✓ 它使用了所有的所谓访问数据的操作 CRUD（创建、读取、更新、删除）。
    - ✓ 当数据存储达到一定量的时候，规模就会几乎保持不变，因为可以从存储中删除过期数据。
  - 数据仓库：某一特定用户发送的邮件；某一特定主题的邮件等。

数据库和数据仓库

- 思考：

✓如何理解数据仓库的特征？

✓数据仓库为什么不需要所有的原始数据？

✓为什么需要细节数据？

✓数据仓库的开发流程

✓数据仓库开发过程中应该注意的问题？

✓如何构建基于数据仓库的DSS？

上海财经大学

SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

THANK YOU