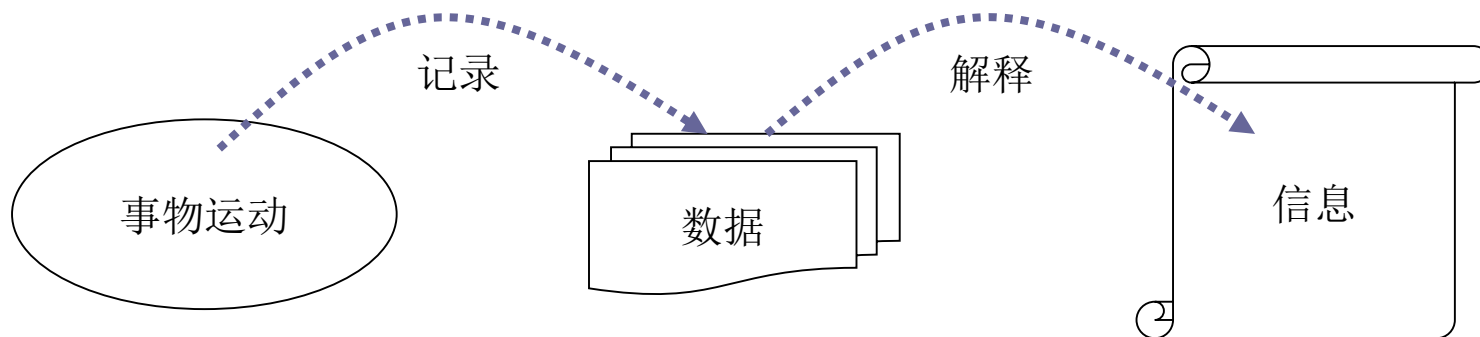


第一章 绪论

- 数据，信息与知识
- 管理就是决策
- 商务智能
- 数据分析
- 数据仓库的基本概念及组成

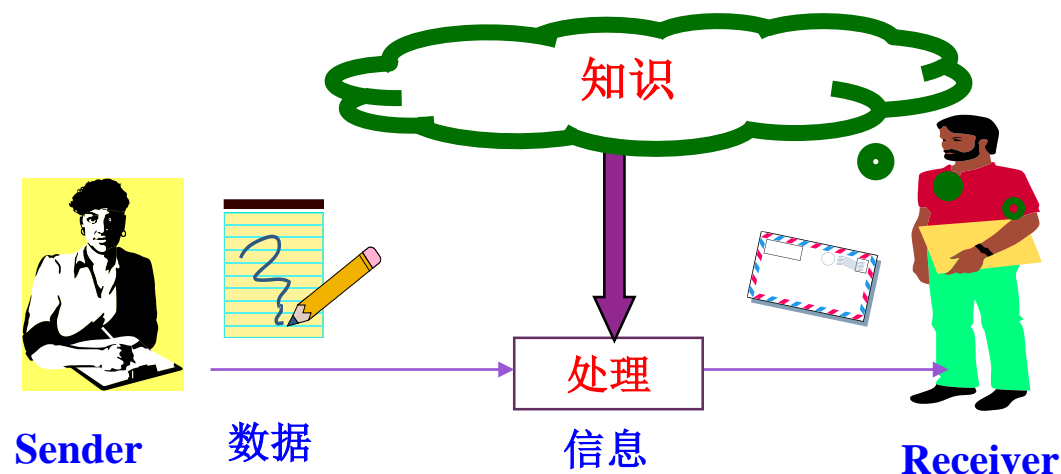
• 数据

- 数据是可以记录、通信和能识别的符号，它通过有意义的组合来表达现实世界中的某种实体（具体对象、事件、状态或活动）的特征。
- 数据类型：
 - 结构化数据、半结构化数据以及非结构化数据
 - 静态的历史数据和动态数据流



• 信息

- 信息是经过某种加工处理后反映客观事物规律的数据
- 数据是信息的载体, **信息是对数据的解释**



• 知识

- 知识是对信息内容进行提炼、比较、挖掘、分析、概括、判断和推论

- **决策**是人们为了达到一定目的而进行的有意识、有选择的行动。
- 决策是企业管理的核心，贯穿管理的过程。**高层决策、中层管理、基层运营**都要决策。
 - 战略层，如厂址选择、资金分配计划、管理体制指定
 - 中间管理层，如销售、财务、生产、人力资源
 - 运营层，也称业务操作层，如“啤酒+尿布”

- 直觉式的决策不一定可靠

- 根据Microsoft公司统计，超过74%的商业决策落后于预定计划或以失败告终，每年损失740多亿美元。
- **案例**：2005年世界第二大零售商家乐福败走日本就是决策失误的结果，日本消费者的消费习惯和欧美的明显不同：欧美国家的许多家庭在周末会驱车到郊区的大型超市采购价格更便宜的食品和用品存放在家里，但日本人的饮食十分讲究新鲜，所以日本的超市一般都设在交通流量大的车站附近或居民比较集中的住宅区，而家乐福在日本开设的超市位于城市的远郊区，没有根据日本不同的商业文化和消费习惯来调整经营策略，导致它在日本“水土不服”。

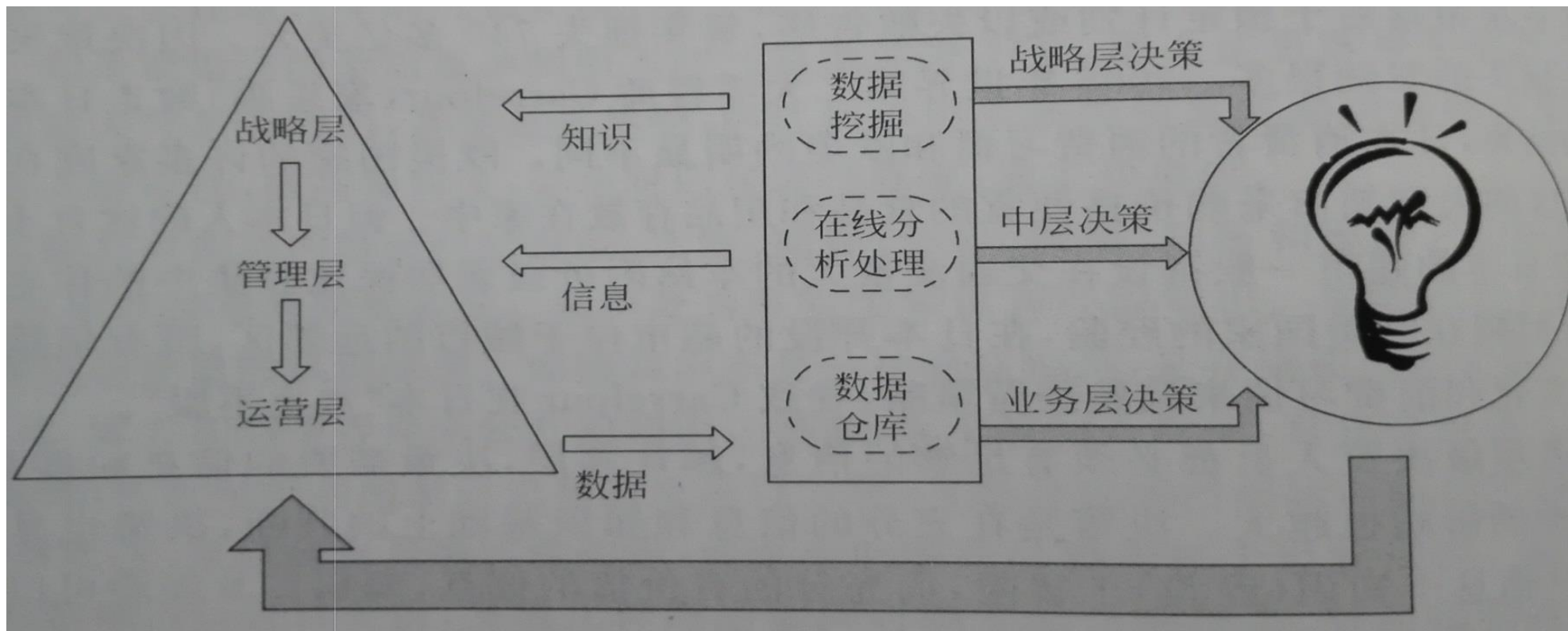
决策需要信息和知识(cont.)



苦恼: 淹没在数据中, 不能制定合适的决策!

- 如何充分利用这些数据资产，挖掘出决策者需要的信息，做出高质量的决策是企业管理者需要考虑的主要问题。
 - **商务智能(Business Intelligence, BI)**把各种数据及时地转换为支持决策的**信息**和**知识**，帮助企业管理者了解顾客的需求、消费习惯、预测市场的变化及行业的整体发展方向，进行有效的**决策**。
- “一个组织获取知识以及把知识快速转化为行动的能力决定其最终的竞争优势” — **前GE CEO Jack Welch**

- 企业界对商务智能有不同的定义
 - **SAP**: 商务智能是收集、存储、分析和访问数据以帮助企业更好决策的技术;
 - **IBM**: 商务智能是一系列技术支持的简化信息收集、分析的策略集合;
 - **Microsoft**: 商务智能是任何尝试获取、分析企业数据以便更清楚了解市场和顾客, 改进企业流程, 更有效地参与竞争的过程;
 - **IDC**: 商务智能是下列软件工具的集合: 终端用户查询和报告工具、在线分析处理工具、数据挖掘软件、数据集市、数据仓库产品和主管信息系统。



- **数据仓库**：用以存储和管理数据，数据从运营层而来。
- **在线分析处理**：用于把这些数据转化为信息，支持各级决策人员复杂查询和在线分析处理，并以直观易懂的图表把结果展现出来。
- **数据挖掘**：从海量数据中提取出隐含在数据中有用的知识供以更有效的决策。

- 什么是数据分析
 - the process of studying the data to find out the answers to **how** and **why** things happened **in the past**. usually, the result of data analysis is a **pattern**, or a **detailed report** that you can further use.
- 数据分析的作用
 - 通过分析可以发现错误，制定新计划以避免重复的错误，让业务变得更好（即便在业务增长的情况下）。

- 数据分析分类

- 描述性分析: what happened

- 以超市为例: 着眼于产品的历史, 发现哪种产品卖得多或哪种产品的需求大, 以此为依据可以在来年进更多的这些产品。

- 诊断性分析: why it happened

- 以超市为例: 如果我们想知道为什么某个产品的需求量较大? 是因为它的品牌还是质量?

- 预测性分析: what will happen

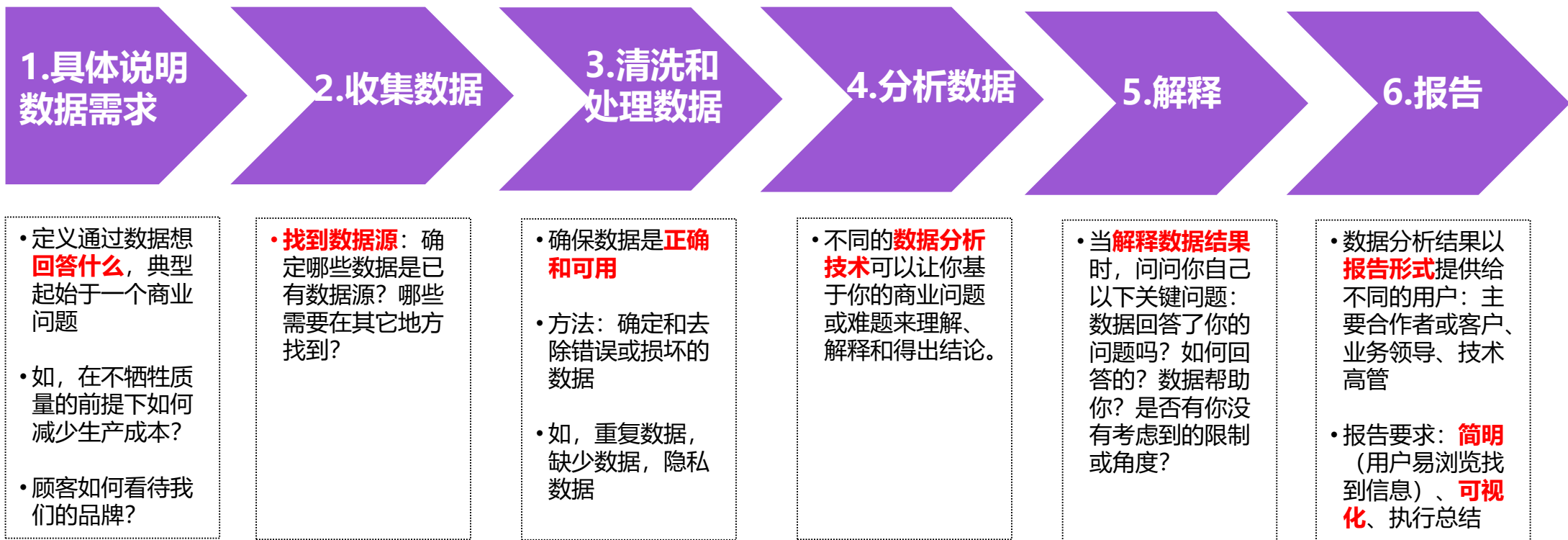
- 预测性分析可以发现将来会发生什么, 通过着眼于过去的趋势和行为模式我们可以预见将来什么会可能发生。
如电商推荐系统

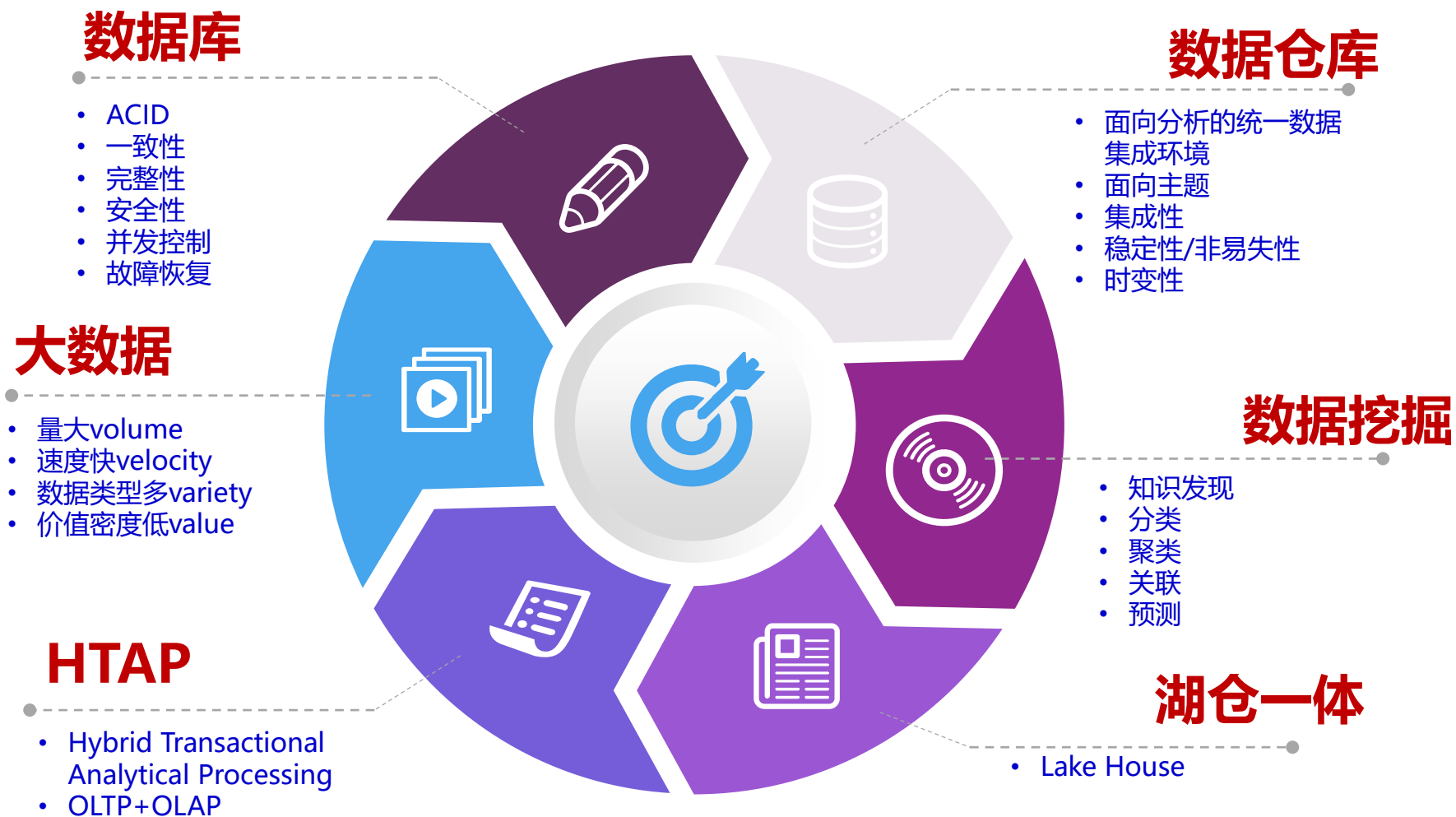
- 指导性分析: how will it happen

- 可以帮助找到哪个是最佳选项。这是一种最高级的分析, 如, 汽车自动驾驶

- 统计分析

- 使用统计方法或技术来分析数据集以便汇总数据中重要和主要的特征, 通常使用一些可视化辅助手段展示。



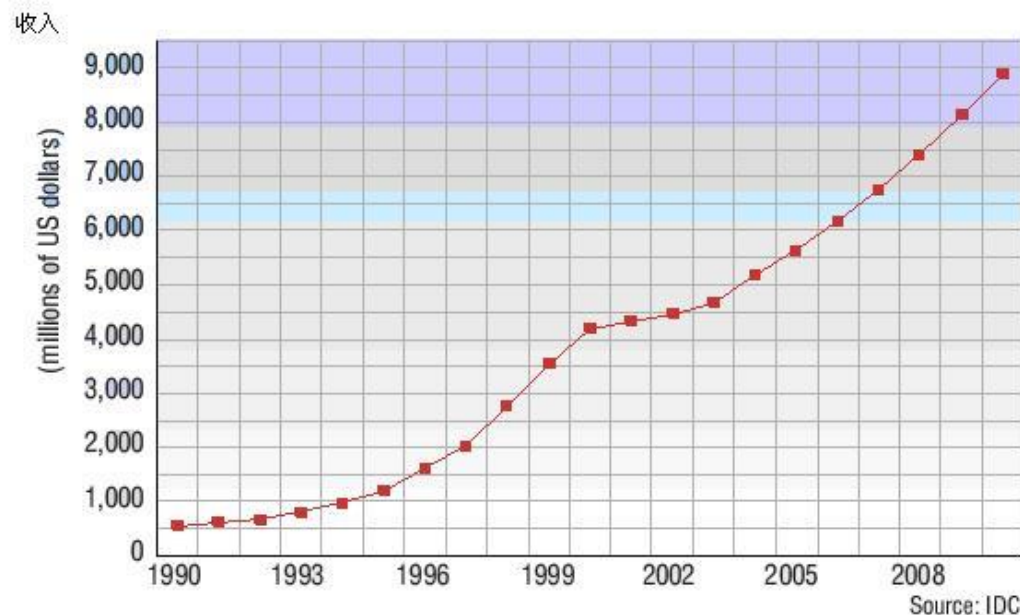


数据仓库的基本概念与组成

- 从数据库到数据仓库
- 数据仓库的定义
- 数据仓库的体系结构
- 数据仓库的数据组织
- 数据源
- ETL
- 元数据
- 数据存储

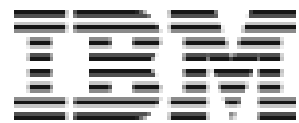
- 数据仓库的发展

- 世界上最早的数据仓库是1981年NCR公司为Wal-Mart 建立的。
- 最早将数据仓库提升到理论高度进行分析并提出数据仓库概念的是W. H. Inmon, Inmon也因此被称为 “数据仓库之父” 。



从数据库到数据仓库(cont.)

• 数据仓库的发展（国际）



- 1998年推出DB2 OLAP服务器
- 2001年收购Informix公司
- 数据库产品线更丰富，成为领导厂商



- SQL Server 7.0



- Sybase IQ
- Warehouse Studio
- SAP data warehouse



- 数据仓库构建、OLAP、数据集市
- Oracle Warehouse Builder



- 开源离线数仓



- AWS Redshift



- Data Warehouse-as-a-Service, 云时代



从数据库到数据仓库(cont.)

• 数据仓库的发展（国内）



- 离线数仓MaxCompute
- 实时数仓Hologres



- 云数据仓库 GaussDB(DWS)



- ByteHouse, 统一的离线和实时数据分析平台

- 事务处理与分析处理

- 数据处理根据管理和使用方式不同可分为两大类：操作型和分析型处理

- 对数据库联机(online)的日常操作，通常是对一个或一组记录的查询或修改，主要为企业的特点应用服务
- 采用实时或在线方式处理数据库
- 关注点：
 - 性能（吞吐量+响应时间）
 - 安全性
 - 完整性

操作型处理特点

- 用于管理人员的决策分析
- 经常要访问大量的历史数据
- 很少的数据库写操作，除非对数据库进行更新或装入时
- 关注点：
 - 数据集成性
 - 数据质量
 - 数据的综合

分析型处理特点

- 事务处理与分析处理不同的具体表现：

- 性能特征不同

- 事务处理环境：用户的行为特点是数据的存取操作频率高，而每次操作处理的时间短。因此系统可以允许多个用户按分时方式使用系统资源，同时保持较短的响应时间。
 - 事务吞吐量使用每秒钟完成的数据处理数TPS/TPM来表示
 - 分析处理环境：用户的行为模式与上面完全不同，一个分析处理程序可能要连续运行几个小时，从而消耗大量系统资源。
 - 在DSS中，吞吐量通常使用每小时处理的查询数QPH来表示
 - 这些查询涉及的数据量非常大，在完成之前通常需要占用绝大多数资源
 - Ad-hoc查询（即席查询）

– 数据集成问题

- **事务处理环境**：目的在于使业务处理自动化，一般只需要与本部门业务相关的当前数据，而对整个企业范围内的集成应用考虑很少，因为大多数企业内部数据的是分散而非集成的。
- **分析处理环境**：需要集成的数据，不仅需要整个企业内部各部门的相关数据，还需要企业外部、竞争对手等处的相关数据。
 - **静态集成**：对所需数据一次性集成，之后就一直以此集成数据做为分析基础，不再与数据源发生联系。它最大缺点在于：当数据源发生变化（集成后），而这些变化不能反映给决策者，导致决策者使用的是过时的数据。
 - **动态集成**：集成数据以一定的周期进行刷新。分析处理需要数据的动态集成。

– 历史数据问题

- **事务处理环境**：一般只需当前数据。数据库中也只存储短期数据，并且不同数据保存期也不相同。即使有历史数据保存，也不利用。
- **分析处理环境**：对决策者而言，历史数据相当重要，许多分析方法必须以大量历史数据为依托，没有对历史数据的详细分析，很难把握企业的发展趋势。

– 数据综合问题

- 事务处理积累了大量的细节数据，一般DSS不对细节数据分析。一是细节数据量大，严重影响分析效率；二是太多的细节数据不利于分析人员将注意力集中在有用信息上。因此，分析处理前经常要综合，而事务处理系统不具备这种综合能力。

从数据库到数据仓库(cont.)

事务型处理数据	分析型处理数据
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	不可更新，只读的
操作需求事先可知道	操作需求事先不知
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
一个时刻操作一个单元	一个时刻操作一组数据
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

- W.H.Inmon(公认的数据仓库定义):
 - 数据仓库是面向主题的、集成的、稳定的、随时间变化的数据集合，用以支持管理决策的过程。
- 数据仓库用来保存从多个数据库或其它信息源选取的数据, 并为上层应用提供统一用户接口，完成数据查询和分析。
- 数据仓库是作为DSS服务基础的分析型DB，用来存放大容量的只读数据，为制定决策提供所需要的信息。
- 数据仓库是与操作型系统相分离的、基于标准企业模型集成的、带有时间属性的、面向主题及不可更新的数据集合。
- 数据仓库是融合方法、技术和工具以在完整的平台上将数据提交给终端用户的一种手段。
- 数据仓库是对分布在企业内部各处的业务数据的整合、加工和分析的过程。

数据仓库的四个特点

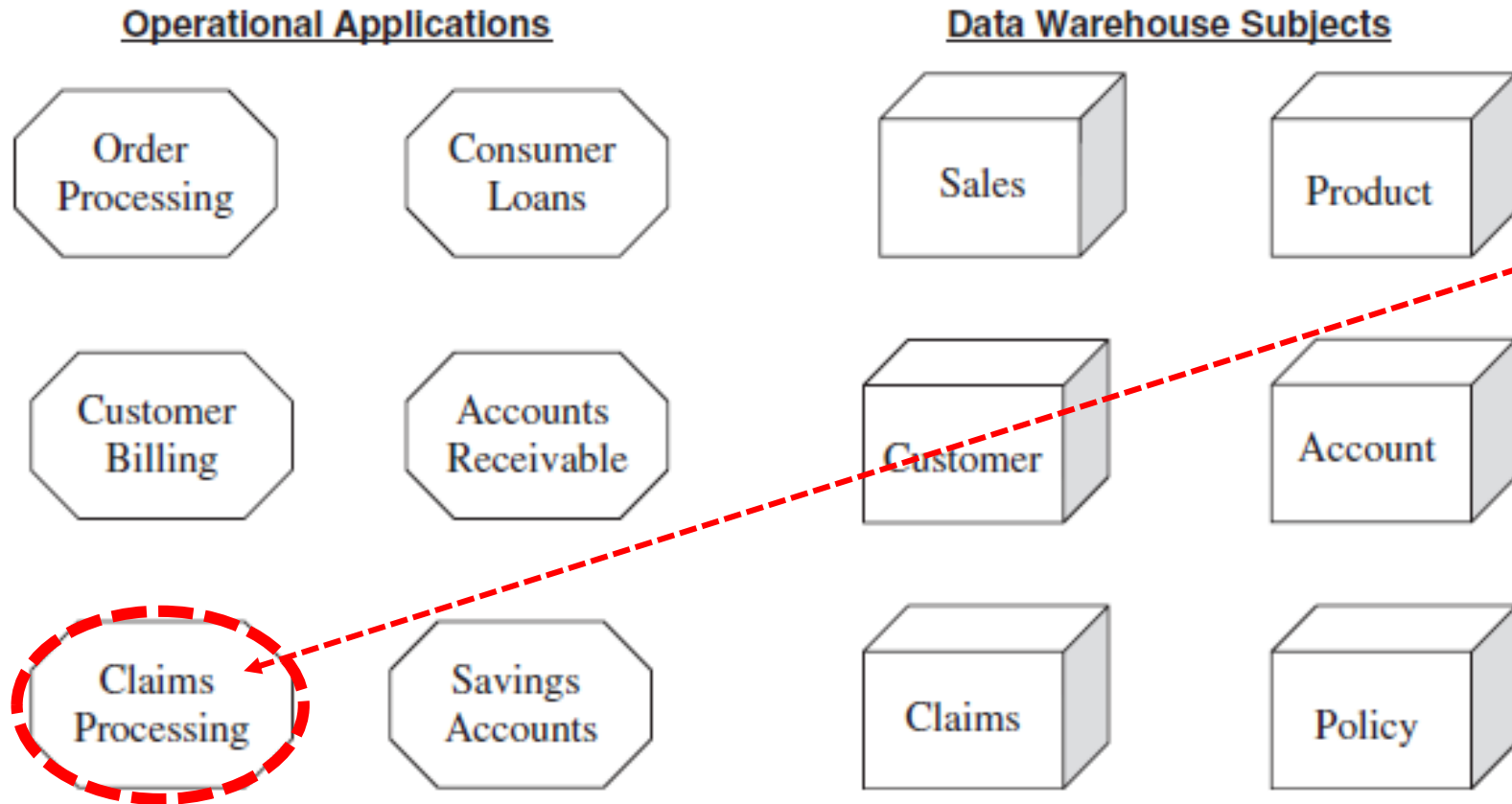
- 数据仓库中的数据有如下特点：
 - 面向主题 (Subject-oriented)
 - 集成的 (Integrated)
 - 稳定的(非易失的) (Non-volatile)
 - 时变的(反映时间变化) (Time-variant)

- 主题subject

- 一个抽象概念，是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
- 逻辑上，它对应于企业中某一宏观分析领域所涉及的分析对象
- 主题随企业的不同而不同
 - 如，对一家制造企业而言，销售、发货和存货都是非常重要的主题，而对一家零售商来说，在付款柜台处的销售就是一个非常重要的主题

特点1.面向主题(cont.)

In the data warehouse, data is not stored by operational applications, but by business subjects.



- 每个应用程序的数据根据应用程序的不同单独组织。
- **索赔**对一家保险公司来说就是非常重要的主题。
- 汽车保险政策的索赔在自动保险应用程序中处理；工人赔偿保险的索赔数据在工人赔偿保险应用程序中。
- 但在保险公司的数据仓库中，索赔数据就按照索赔的主题进行组织

Figure 2-1 The data warehouse is subject oriented.

特点1.面向主题(cont.)

- 面向主题

- 面向主题的数据组织方式可在较高层次上对分析对象的数据给出完整、一致的描述，能完整、统一的刻画各个分析对象所涉及企业的各项数据以及数据之间的联系，从而适应企业各个部门的业务活动特点和企业数据的动态特征，从根本上实现数据与应用的分离
- 面向应用的数据经常会随着各种经营环境的改变而发生变化，面向主题的数据则因为比应用具有更高的抽象层次而比较稳定
- 但数据的产生都是基于应用而产生，因此数据在进入数据仓库之前，必然要经过加工和集成，将原始数据做一个从面向应用到面向主题的大转变

特点1.面向主题(cont.)

• 示例:

- 现有一家采用“会员制”经营方式的商场，按业务建立起若干子系统，并按业务处理要求建立各自数据库模式

采购子系统: 订单(订单号, 供应商号, 总金额, 日期) 订单细节(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, 地址, 电话) 销售(员工号, 顾客号, 商品号, 单价, 数量, 日期)
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, 部门主管, 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)

特点1.面向主题(cont.)

- 面向应用到面向主题的转变
 - 面向主题的数据组织方式应分为两个步骤：
 1. 抽取主题
 2. 确定每个主题所包含的数据内容

仍以商场为例，它所应有的主题包括：商品、供应商、顾客。每个主题有各自独立的逻辑内涵，对应一个分析对象。

特点1.面向主题(cont.)

- **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

- **供应商：**

- 供应商固有信息：商品号，商品名，类别，颜色等
- 供应商商品信息：供应商号，供应价，供应日期，供应量等

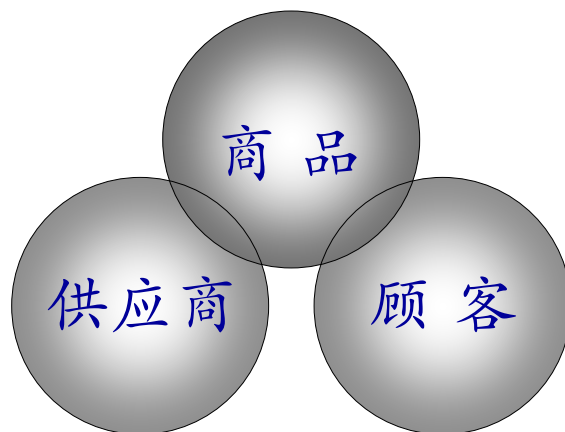
- **顾客：**

- 顾客固有信息：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

特点1.面向主题(cont.)

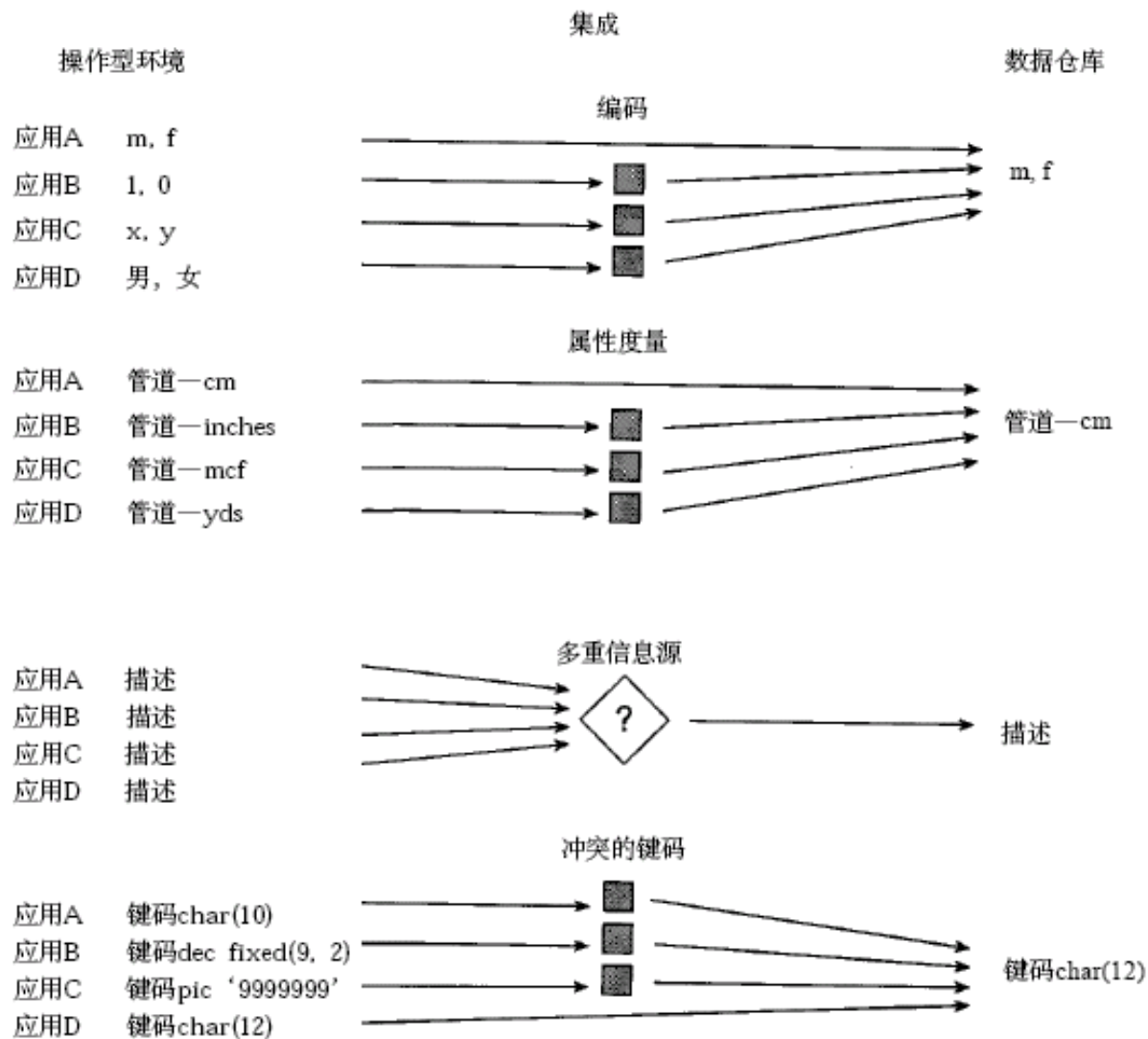
• 从面向应用 ⇨ 面向主题

- 丢弃了原来不必要，不适合分析的信息
- 将分散在各子系统中的有关主题的信息集成，形成关于主题的一致信息
- 不同主题之间也有重叠的内容，但只是逻辑上的重叠，细节级上的重叠，另外主题间并不是两两重叠



- 数据仓库中的数据是从原有分散的源数据库中提取出来的，其每一个主题所对应的源数据在原有的数据库中有许多冗余和不一致，且与不同的应用逻辑相关
- 为了创建一个有效的主题域，必须将这些来自不同数据源的数据集成起来，使之遵循统一的编码规则
- 主要两个工作：统一源数据所有矛盾之处；进行数据综合和计算
 - 数据仓库在提取数据时必须经过数据集成，消除源数据中的矛盾，并进行数据综合和计算。经过数据集成后，数据仓库所提供的信息比数据库提供的信息更概括、更本质

特点2.集成性(cont.)



特点2.集成性(cont.)

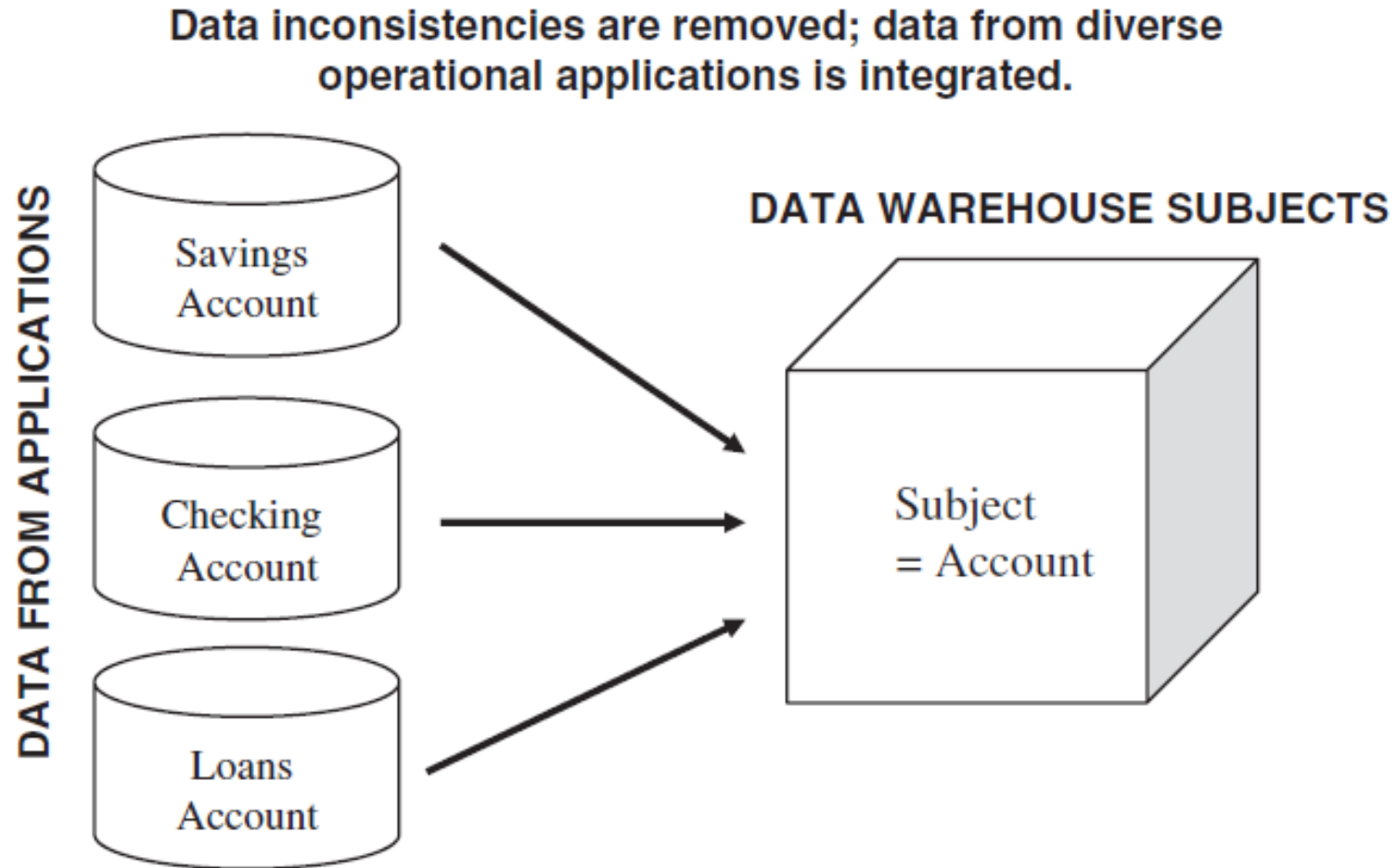


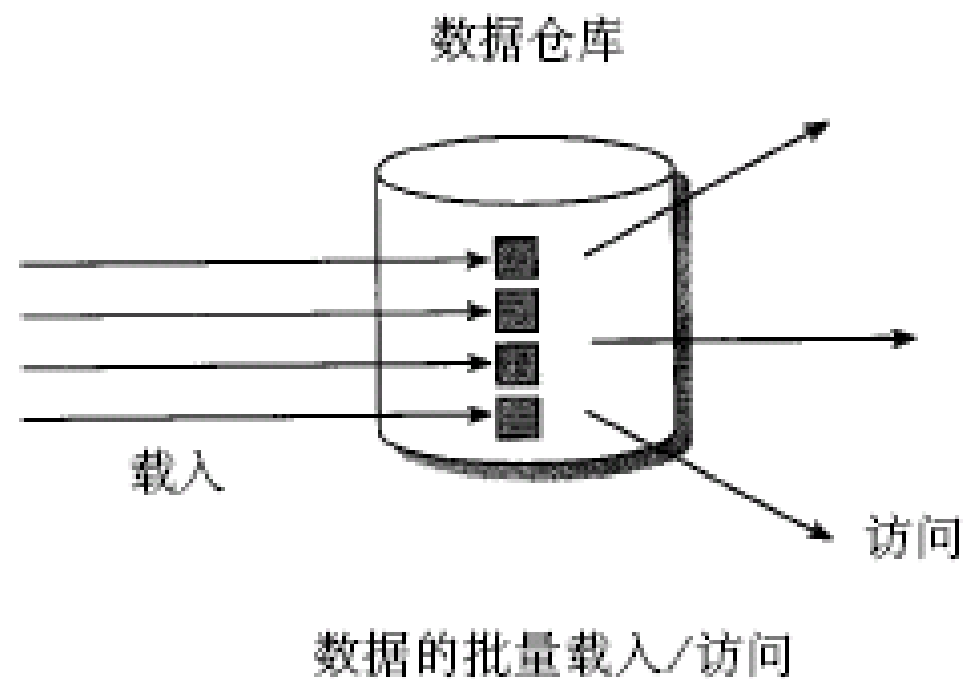
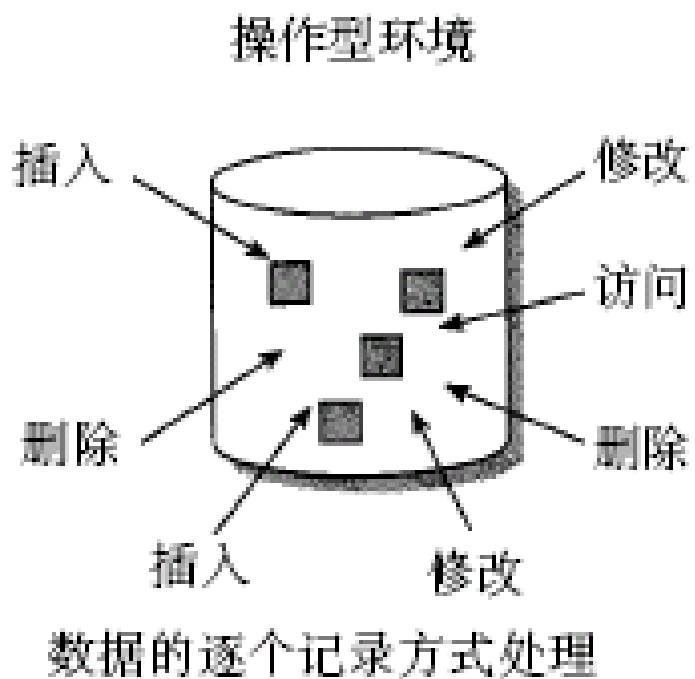
Figure 2-2 The data warehouse is integrated.

- 从操作型系统中提取的数据和从外部数据源中取得的数据，每隔一段**时间**被存储到数据仓库中。
 - 根据商业交易的需要，这种过程一般来说一天两次，一天一次，一个星期一次，或两个星期一次都是可以的。
- 在每次商业交易发生时，可以**实时地更新操作型系统中的数据**，但并不频繁地对数据仓库进行更新。
- 不能在数据仓库中实时地删除数据。**一旦数据存入了数据仓库，就不能对这个数据进行修改**。因为数据仓库中的数据是用来查询和分析的

特点3.稳定性(cont.)

- 数据仓库中的数据反映的是一段时间内历史数据的内容，是不同时间点的**数据库快照的集合**，以及基于撰写快照进行统计、综合和重组的导出数据，而不是联机处理的数据。
- 主要供企业高层决策分析之用，所涉及的数据操作主要是**查询**，一般情况下**并不进行修改操作**，即数据仓库中的数据是不可实时更新的，仅当超过规定的存储期限，才将其从数据仓库中删除，提取新的数据经集成后输入数据仓库。

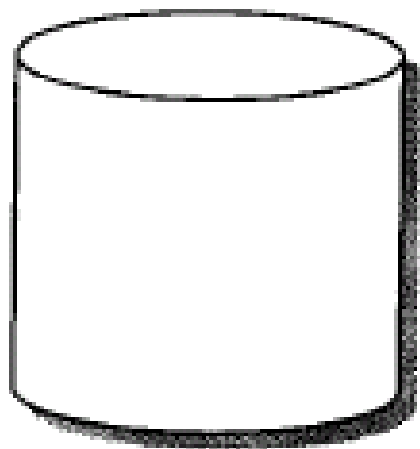
特点3.稳定性(cont.)



- 许多商业分析要求对**发展趋势做出预测**，对发展趋势的分析需要访问历史数据
- 因此数据仓库必须**不断捕捉**事务数据库中变化的数据，生成数据库的快照，经集成后增加到数据仓库中去
- 另外数据仓库还需要随时间的变化删去过期的、对分析没有帮助的数据，且需要按规定的时间段增加综合数据

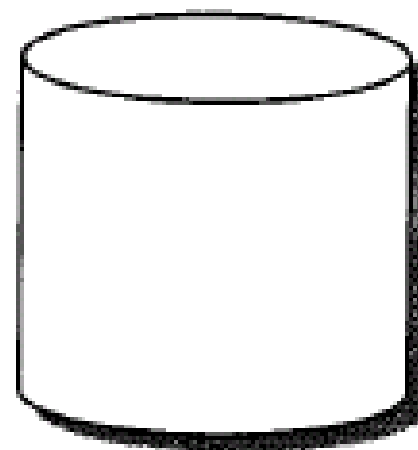
特点4.时变(cont.)

操作型环境



- 时间期限：当前到60~90天
- 记录更新
- 键码结构可能包括/也可能不包括时间元素

数据仓库

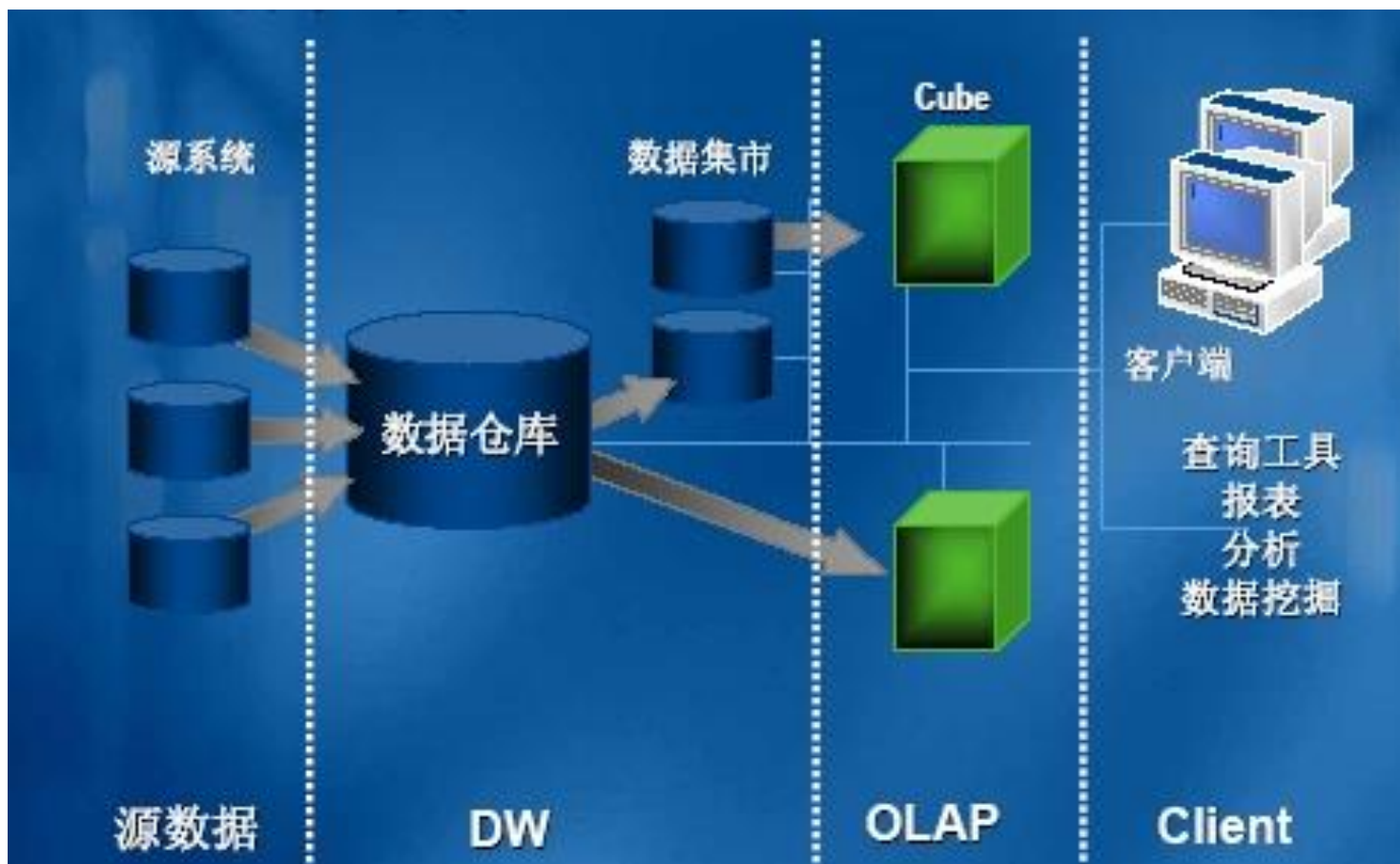


- 时间期限：5~10年
- 数据的复杂快照
- 键码结构包括时间元素

- 数据仓库支持OLAP、数据挖掘和决策分析
- OLAP从数据仓库中的综合数据出发，提供面向分析的**多维模型**，并使用**多维分析**的方法从**多个角度**、**多个层次**对**多维数据**进行分析，使决策者能够以更加自然的方式来分析数据
- 数据挖掘则以数据仓库和多维数据库中的数据为基础，发现数据中的**潜在模式**和**进行预测**
- 因此，数据仓库的功能是支持管理层进行科学决策，而**不是事务处理**

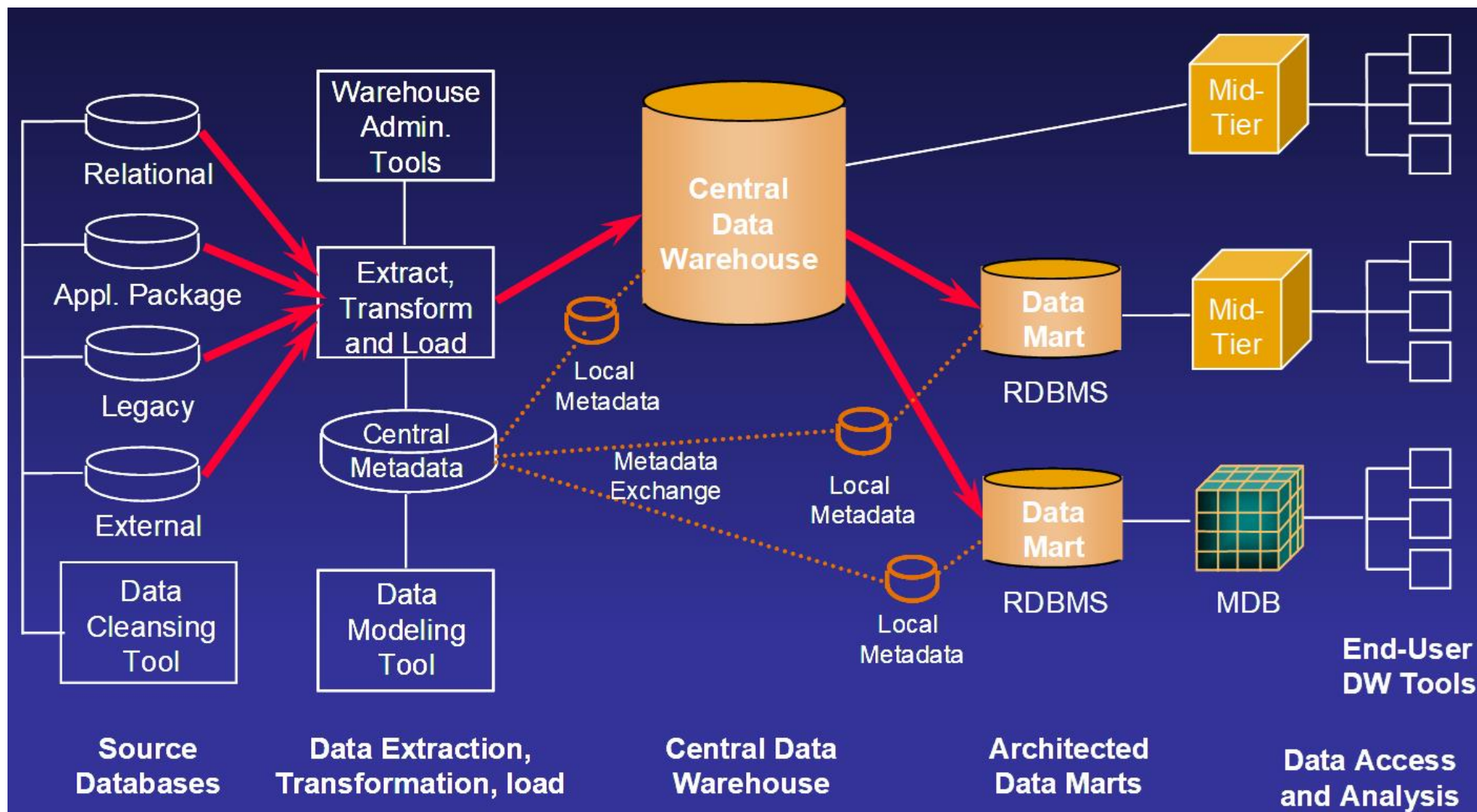
- 大量数据的组织和管理
 - 包含了大量的历史数据，它是从数据库中提取得来的，如何高效的组织和管理数据是数据仓库性能能否发挥的重要前提。
- 复杂分析的高性能体现
 - 涉及大量数据的聚集、综合等，在进行复杂查询时经常会使用多表的联接、累计、分类、排序等操作。
- 对提取出来的数据进行集成
 - 数据仓库中的数据是从多个应用领域中提取出来的，在不同的应用领域和不同的数据库系统中都有不同的结构和形式，所以如何对数据进行集成也是构建数据仓库的一个重要方面。
- 对进行高层决策的最终用户的界面支持
 - 提供各种分析应用工具

数据仓库的体系结构

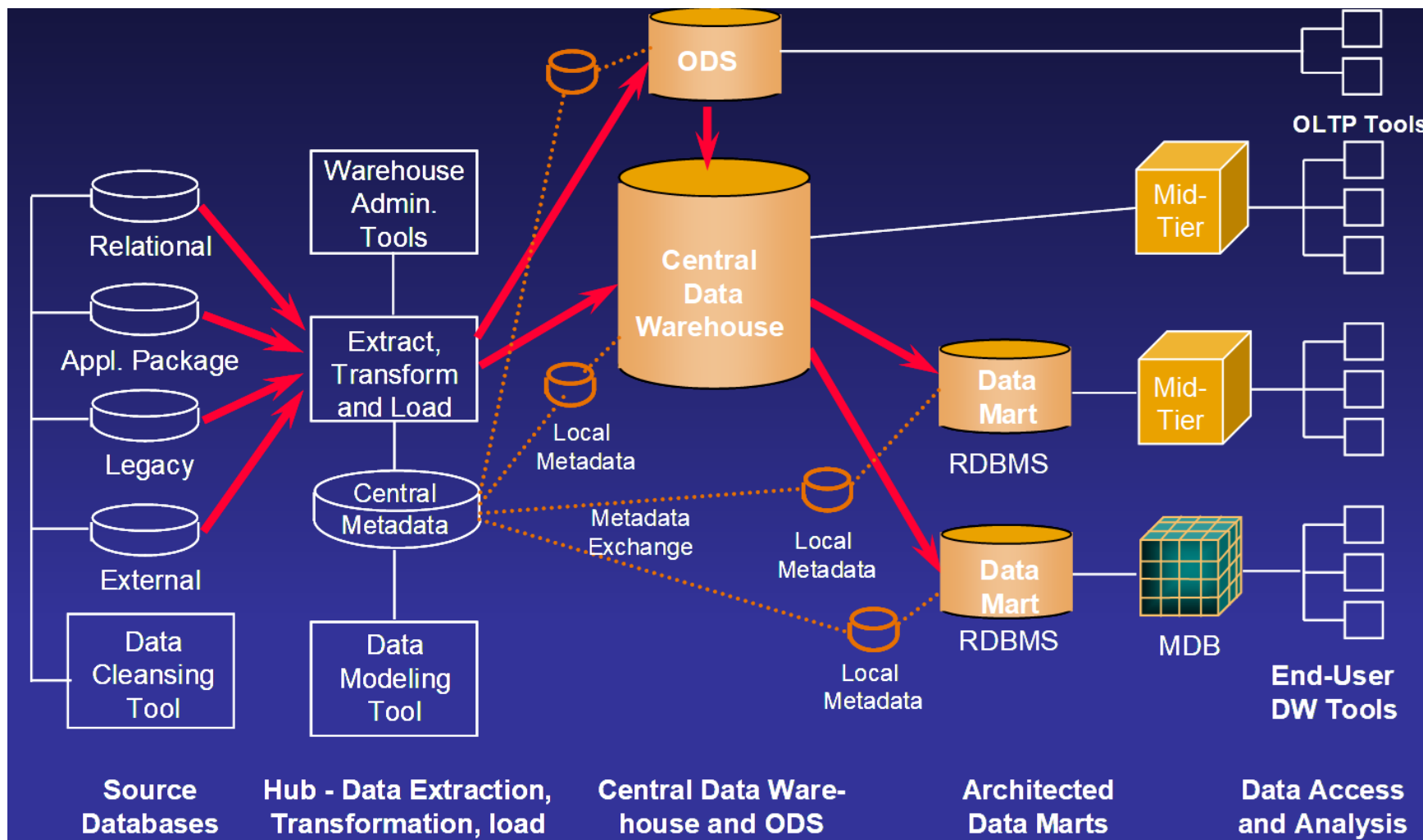


传统通用数据仓库体系结构

数据仓库的体系结构(cont.)



数据仓库的体系结构(cont.)



帶ODS的数据仓库体系结构

- ETL
- 数据仓库存储
- 元数据
- 数据集市
- OLAP

- Extract, Transform, Load的缩写
- ETL是从数据源中抽取数据并转换到数据仓库中的过程，包括四个过程：
 - 数据抽取(Data Extraction)
 - 数据仓库按照主题组织数据，只抽取系统分析需要的数据
 - 数据转换(Data Transformation)
 - 业务系统可能采用不同的数据库产品（Oracle,DB2, Sybase等），数据类型可能不同，需要转换。
 - 数据清洗(Data Cleaning)
 - 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，以免影响系统决策的正确性。
 - 数据装载(Data Loading)
 - 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

- 数据存储

- 用于存放数据仓库数据和元数据的存储空间
- 存储方式有三种：多维数据库、关系型数据库以及前两种存储方式的结合

- 元数据

- 描述数据的数据
- 管理元数据
 - 数据仓库设计人员使用的描述数据仓库的数据信息，用于执行数据仓库开发和管理任务
- 用户元数据是帮助用户查询信息、理解结果及了解数据仓库中的数据和组织

元数据管理工具Atlas, <https://blog.csdn.net/u012543380/article/details/110070153>

- 数据集市(Data mart)
 - 面向企业中某个部门(主题)而在逻辑上或物理上划分出来的数据仓库的一个数据子集。
- OLAP
 - Online Analytical Processing联机分析处理
 - 是使分析人员、管理人员或执行人员能够从多角度对信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术
 - 目的是满足决策支持或者满足在多维环境下特定的查询和报表需求，技术核心是“维”的概念
 - 包括：切片、切块、钻取、旋转等各种分析动作

数据仓库的数据组织

- 数据粒度
- 数据分割

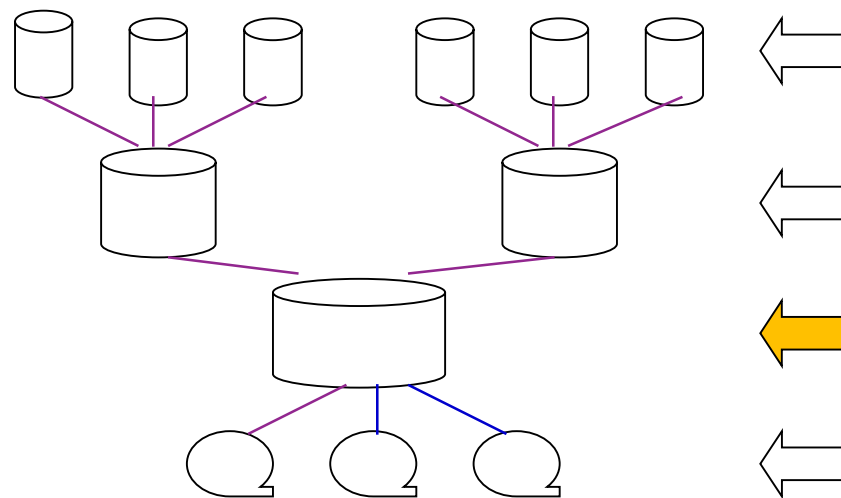
- **粒度**(Granularity)是指数据仓库中数据的**综合级别**

2013-2023月销售

2013-2023周销售

2013-2023销售细节

2002-2012销售细节



高度综合级

轻度综合级

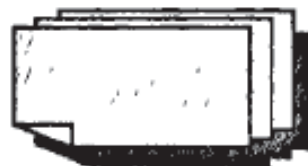
当前细节级

早期细节级

多级数据组织结构

数据粒度(cont.)

高细节级(低粒度级)



例如: 一个顾客一个月的
每个电话细节



每月40000个字节
每月200条记录

01 通话记录

02 日期

02 时间

02 被叫号码

02 接线员帮助

02 完成通话

02 通话时长

02 长途

02 本地网

02 特殊费用

.....

.....

低细节级(高粒度级)



例如: 一个顾客一个月
的电话综合



每月200个字节
每月1条记录

01 通话记录

02 月份

02 电话次数

02 电话平均时长

02 长途电话次数

02 本地通话次数

.....

.....

- 第一种形式：综合程度

- 对数据仓库中的数据综合程度高低的一个度量，它既影响数据仓库中的数据量的多少，也影响数据仓库所能回答询问的种类。
- 粒度越小，综合程度越低，回答查询的种类越多；粒度越高，综合程度越高，查询的效率也越高。
- 在数据仓库中可将小粒度的数据存储在低速存储器上；大粒度的数据存储在高速存储器上。

粒度的两种不同表现形式(cont.)

- 第二种形式：样本数据库

- 样本数据库：在分析过程中，有许多探索的过程有时分析的目的并不要求精确的结果，只需要得到相对准确、能反映趋势的数据，所以可以提取出样本数据库
- 样本数据库的粒度：是根据采样率的高低来划分的，采样粒度不同的样本数据库可以具有相同的综合级别，它是按一定的采样率从细节数据库或轻度综合数据库中提取的一个子集
- 样本数据库的抽取按照数据的重要程度不同进行，利用样本数据库采集重要数据进行分析既可提高分析效率，又有助于抓住主要因素和主要矛盾