第一章 绪论

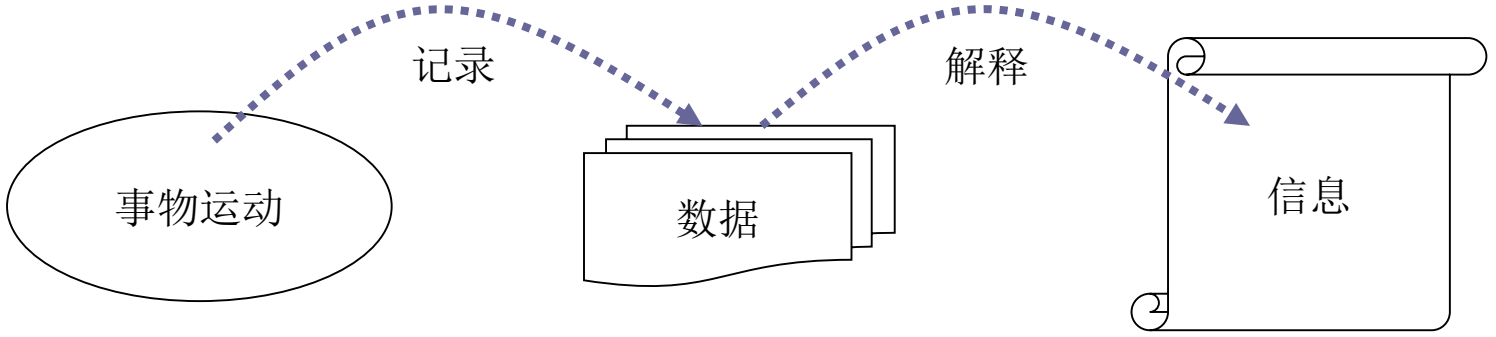
本章大纲

厦門大學 XIAMEN UNIVERSITY

- 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

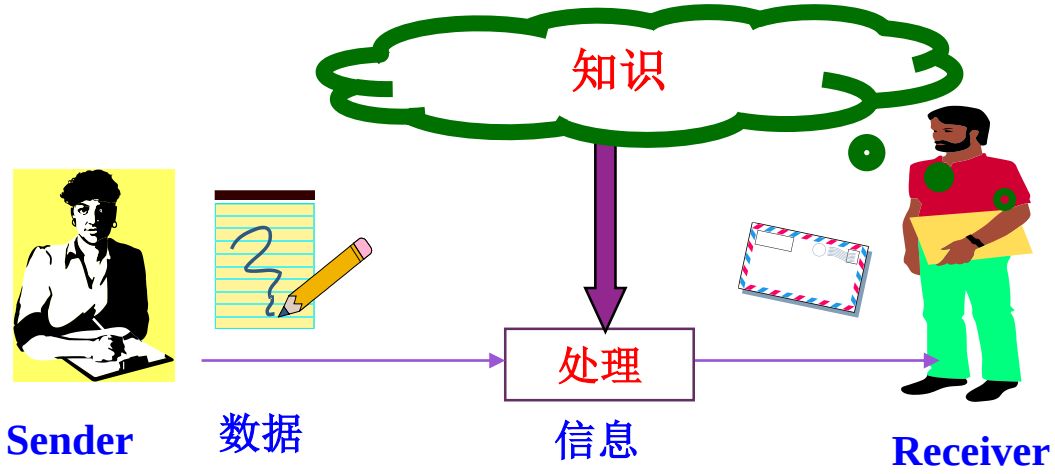
数据，信息与知识

- 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



数据，信息与知识(cont.)

- 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

管理就是决策

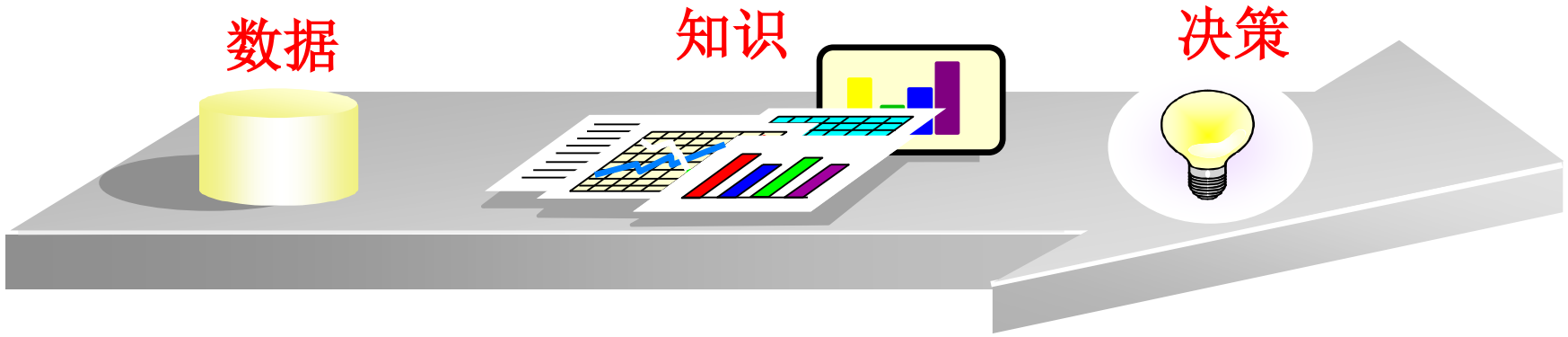
- 决策是人们为了达到一定目的而进行的有意识、有选择的行动。

- 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”

决策需要信息和知识

- 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.)



数据	知识	决策
金融	模式	目标市场
经济	趋势	资金分配
POS	事实	贸易选择
人口统计	关系	在哪儿做广告
生命周期	模型	销售的地理位置
	关联规则	
	序列	

苦恼: 淹没在数据中，不能制定合适的决策!

商务智能

- 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**。

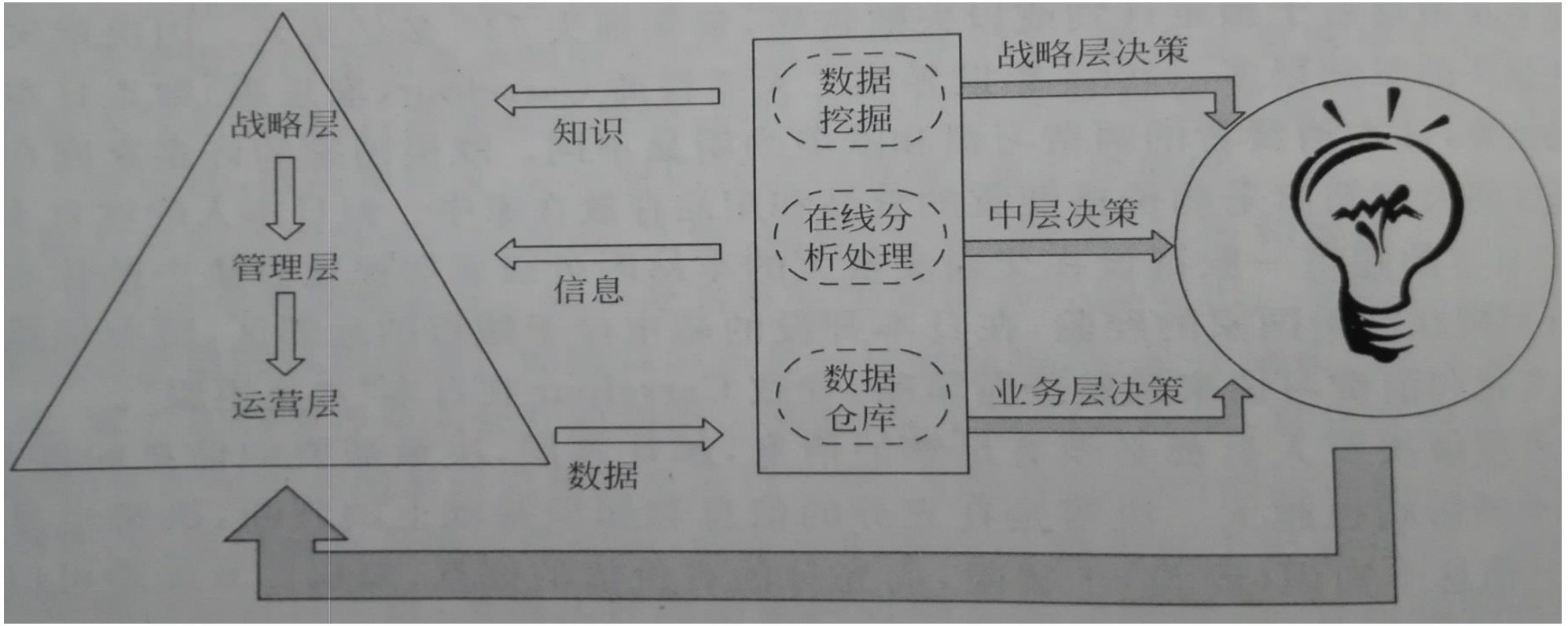
- “一个组织获取知识以及把知识快速转化为行动的能力决定其最终的竞争优势”—前GE CEO Jack Welch

商务智能(cont.)

厦門大學 XIAMEN UNIVERSITY

- 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

商务智能(cont.)



- **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。

数据分析

厦門大學 XIAMEN UNIVERSITY

- 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(cont.)

厦門大學 XIAMEN UNIVERSITY

- 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。

数据分析过程

厦門大學 XIAMEN UNIVERSITY

1.具体说明数据需求

- 定义通过数据想**回答什么**，典型起始于一个商业问题
- 如，在不牺牲质量的前提下如何减少生产成本？
- 顾客如何看待我们的品牌？

2.收集数据

- **找到数据源**：确定哪些数据是已有数据源？哪些需要在其它地方找到？

3.清洗和处理数据

- 确保数据是**正确和可用**
- 方法：确定和去除错误或损坏的数据
- 如，重复数据，缺少数据，隐私数据

4.分析数据

- 不同的**数据分析技术**可以让你基于你的商业问题或难题来理解、解释和得出结论。

5.解释

- 当**解释数据结果**时，问问你自己以下关键问题：数据回答了你的问题吗？如何回答的？数据帮助你？是否有你没有考虑到的限制或角度？

6.报告

- 数据分析结果以**报告形式**提供给不同的用户：主要合作者或客户、业务领导、技术高管
- 报告要求：**简明**（用户易浏览找到信息）、**可视化**、执行总结

数据分析相关技术

数据库

- ACID
- 一致性
- 完整性
- 安全性
- 并发控制
- 故障恢复

数据仓库

- 面向分析的统一数据集成环境
- 面向主题
- 集成性
- 稳定性/非易失性
- 时变性

大数据

- 量大volume
- 速度快velocity
- 数据类型多variety
- 价值密度低value

数据挖掘

- 知识发现
- 分类
- 聚类
- 关联
- 预测

HTAP

- Hybrid Transactional Analytical Processing
- OLTP+OLAP

湖仓一体

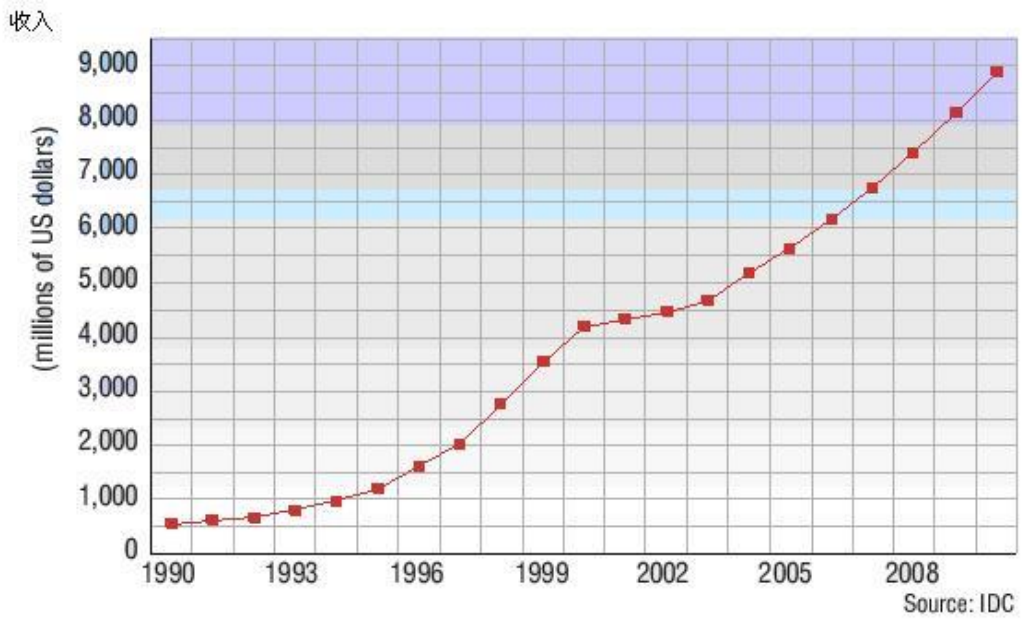
- Lake House

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

从数据库到数据仓库

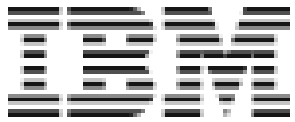
- 数据仓库的发展
  - 世界上最早的数据仓库是1981年NCR公司为Wal-Mart 建立的。
  - 最早将数据仓库提升到理论高度进行分析并提出数据仓库概念的是W. H. Inmon，Inmon也因此被称为“数据仓库之父”。



从数据库到数据仓库(cont.)

厦門大學 XIAMEN UNIVERSITY

- 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

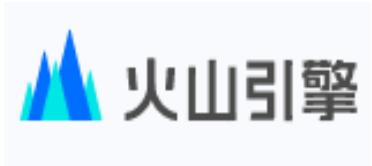
- Data Warehouse-as-a-Service,云时代

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

从数据库到数据仓库(cont.)

厦門大學 XIAMEN UNIVERSITY

- 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

操作型处理特点	分析型处理特点
• 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	• 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

从数据库到数据仓库(cont.)

厦門大學 XIAMEN UNIVERSITY

- 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）

从数据库到数据仓库(cont.)

厦門大學 XIAMEN UNIVERSITY

– 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。

从数据库到数据仓库(cont.)

- 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.)

厦門大學 XIAMEN UNIVERSITY

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

数据仓库的定义

- 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

厦門大學 XIAMEN UNIVERSITY

- 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）

特点1.面向主题

厦門大學 XIAMEN UNIVERSITY

- 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.)

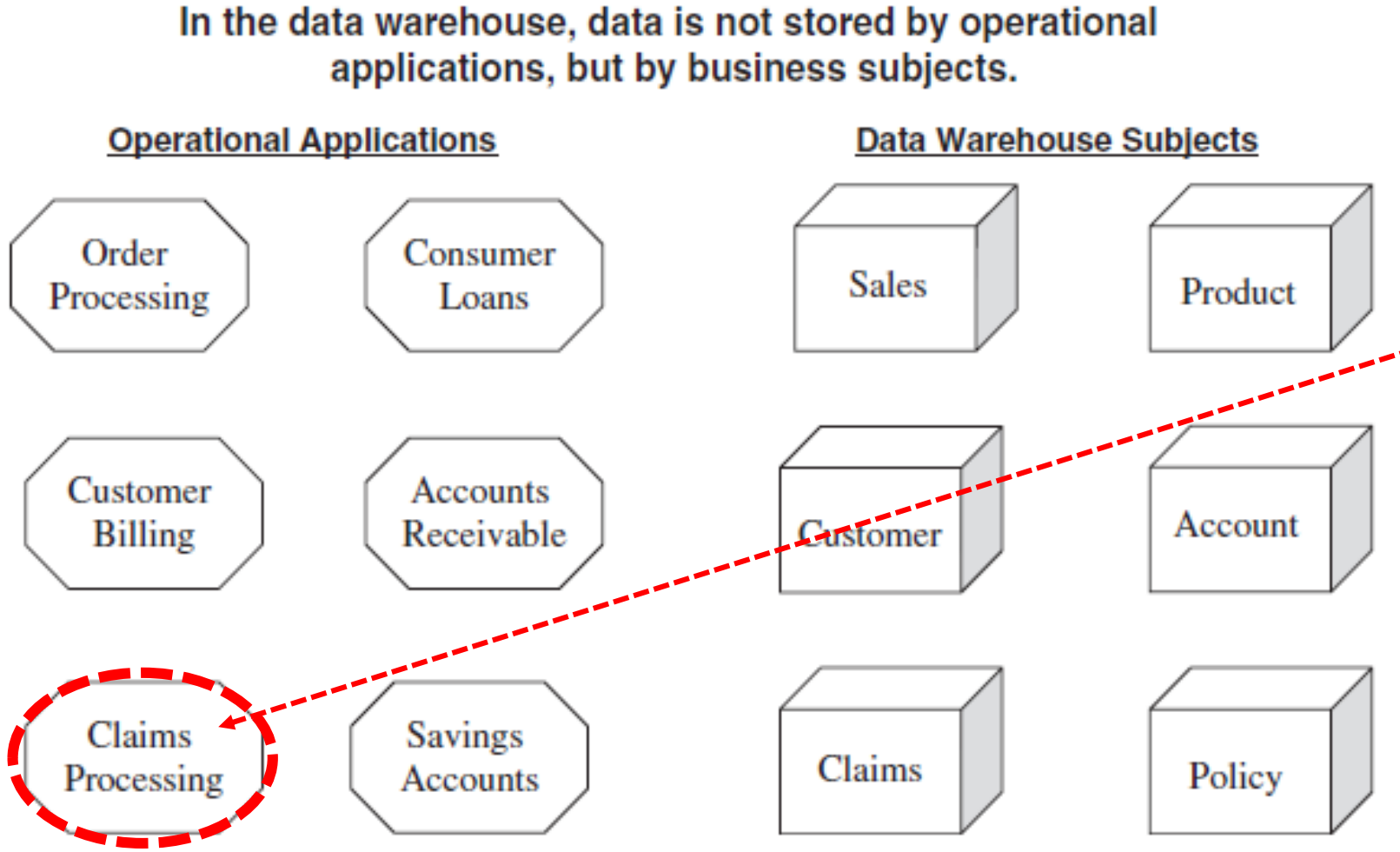


Figure 2-1 The data warehouse is subject oriented.

- 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

特点1.面向主题(cont.)

厦門大學 XIAMEN UNIVERSITY

- 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.)

厦門大學 XIAMEN UNIVERSITY

- **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.)

厦門大學 XIAMEN UNIVERSITY

- 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。

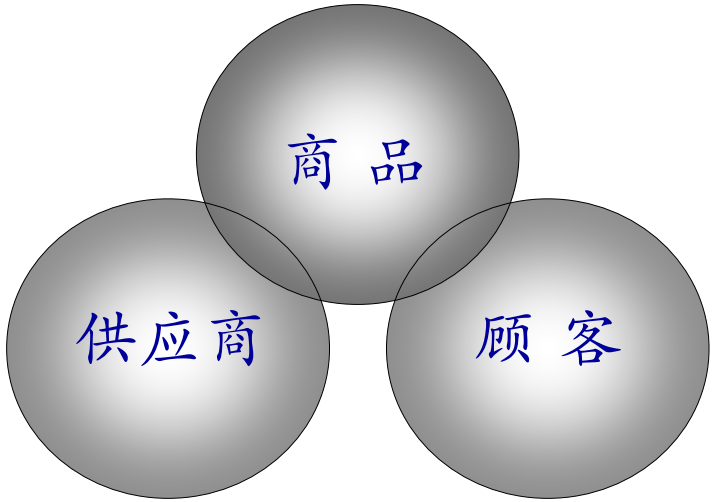
特点1.面向主题(cont.)

厦門大學 XIAMEN UNIVERSITY

- **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



特点2.集成性

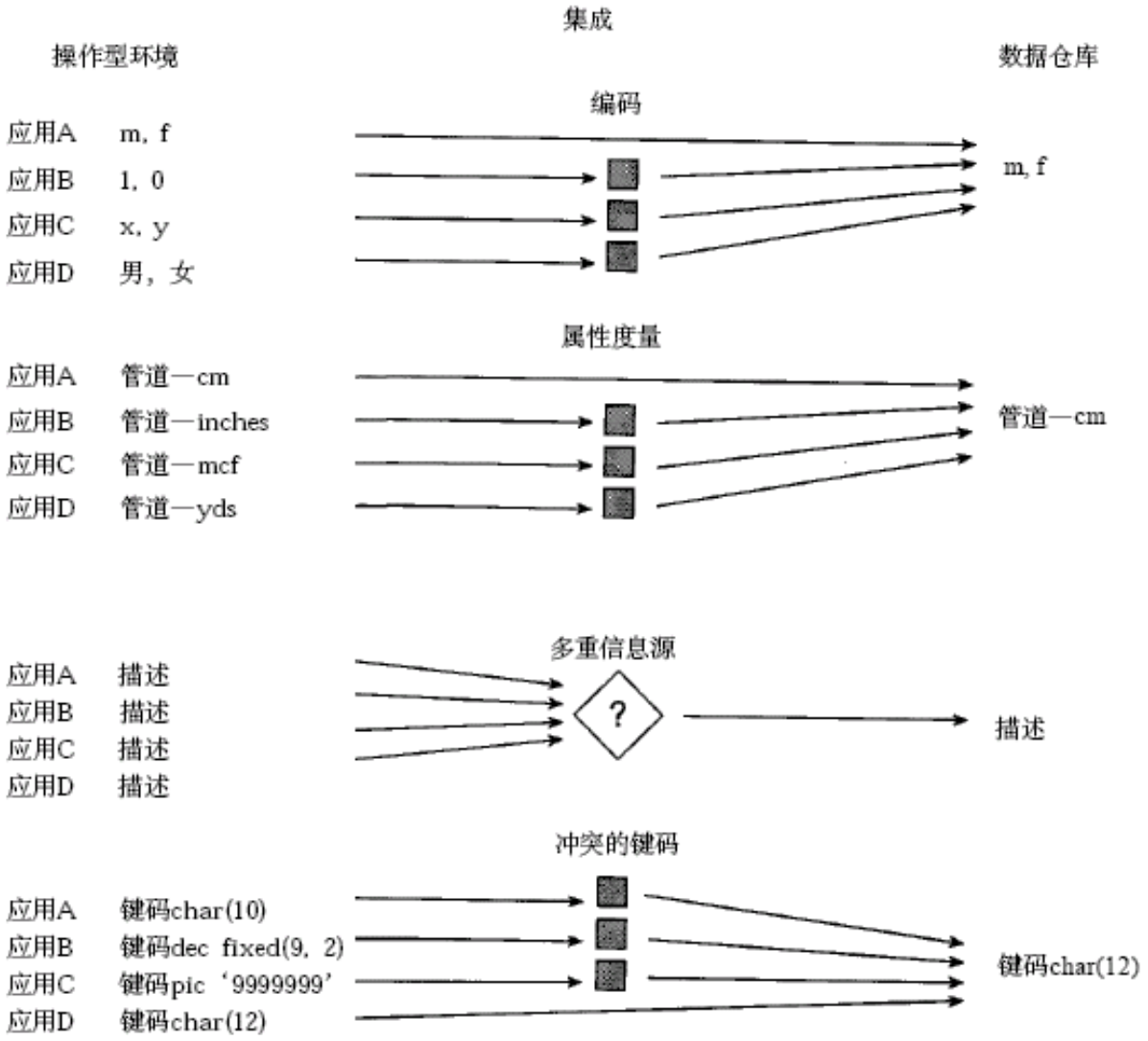
厦門大學 XIAMEN UNIVERSITY

- 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.)


集成
操作型环境
数据仓库
编码
应用A m, f
应用B 1, 0
应用C x, y
应用D 男，女
m, f
属性度量
应用A 管道—cm
应用B 管道—inches
应用C 管道—mcf
应用D 管道—yds
管道—cm
多重信息源
应用A 描述
应用B 描述
应用C 描述
应用D 描述
?
描述
冲突的键码
应用A 键码char(10)
应用B 键码dec fixed(9, 2)
应用C 键码pic '9999999'
应用D 键码char(12)
键码char(12)

特点2.集成性(cont.)

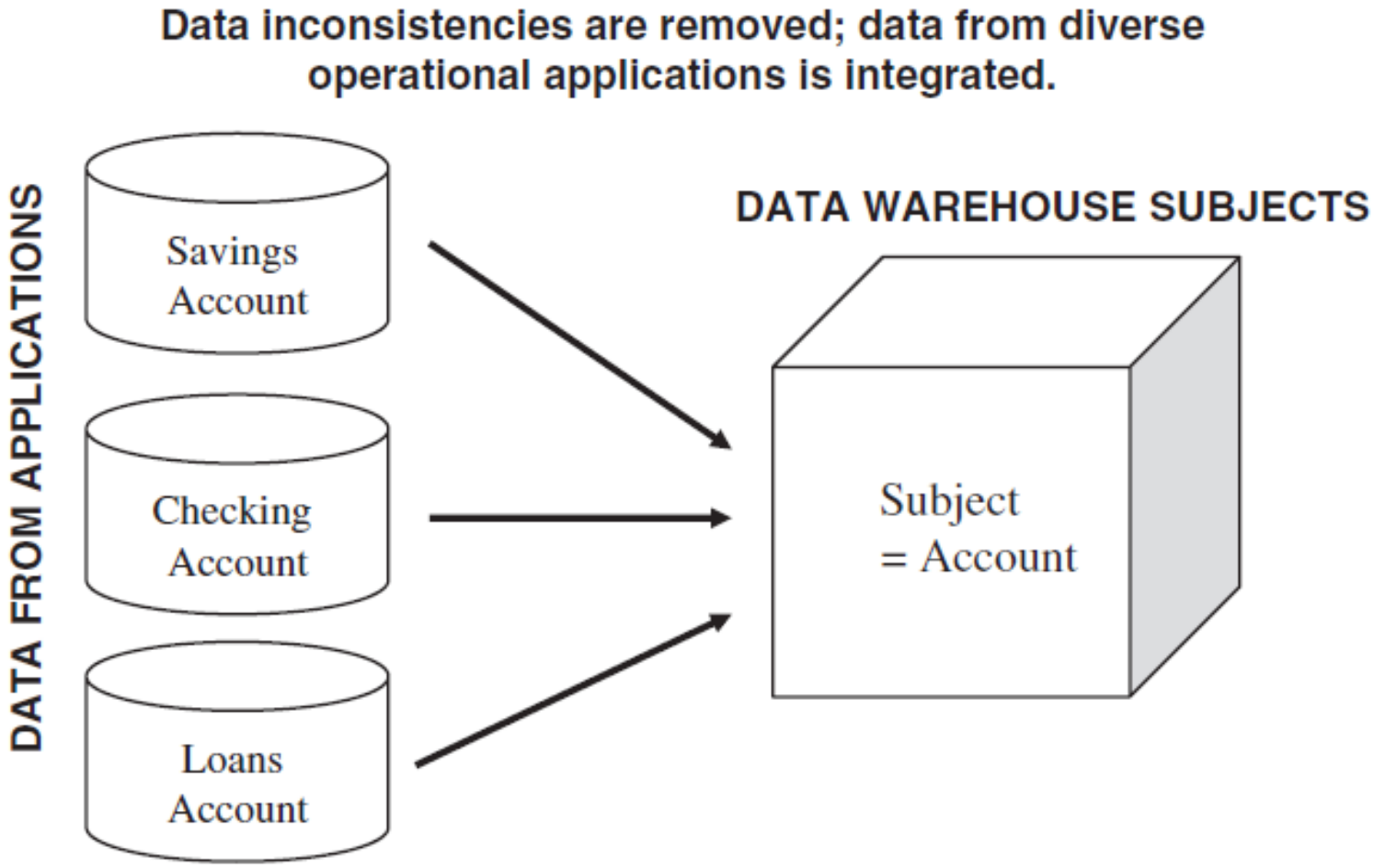


Figure 2-2 The data warehouse is integrated.

特点3.稳定性

- 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。

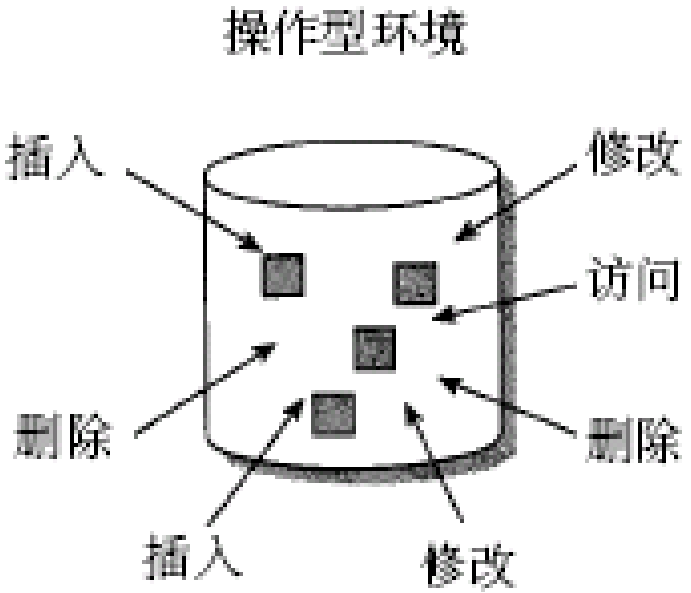
- 不能在数据仓库中实时地删除数据。一旦数据存入了数据仓库，就不能对这个数据进行修改。因为数据仓库中的数据是用来查询和分析的

特点3.稳定性(cont.)

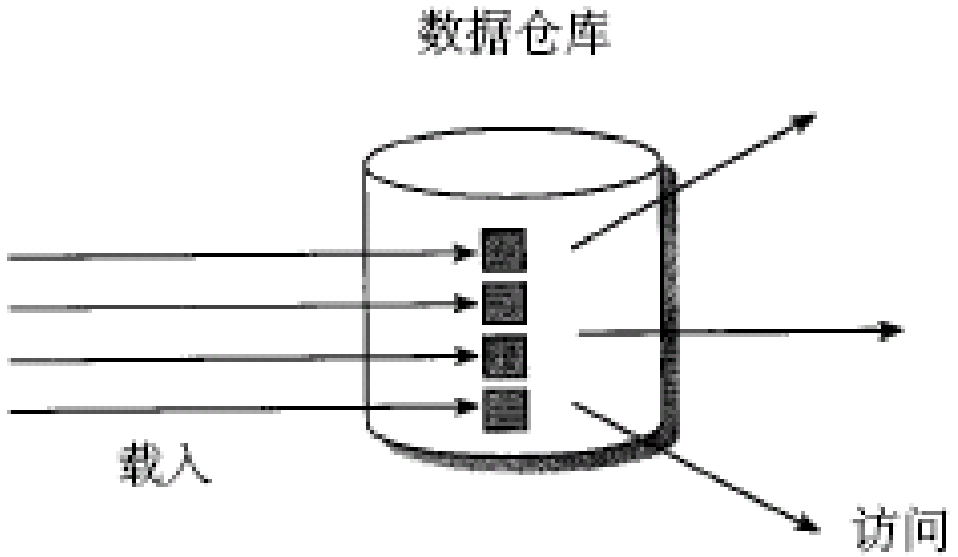
- 数据仓库中的数据反映的是一段时间内历史数据的内容，是不同时间点的数据库快照的集合，以及基于撰写快照进行统计、综合和重组的导出数据，而不是联机处理的数据。

- 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.)



数据的逐个记录方式处理



数据的批量载入/访问

特点4.时变

厦門大學 XIAMEN UNIVERSITY

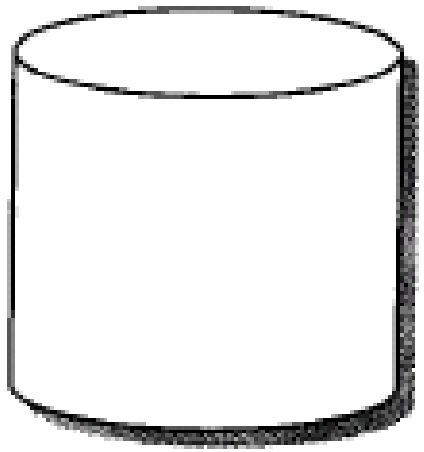
- 许多商业分析要求对发展趋势做出预测，对发展趋势的分析需要访问历史数据

- 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

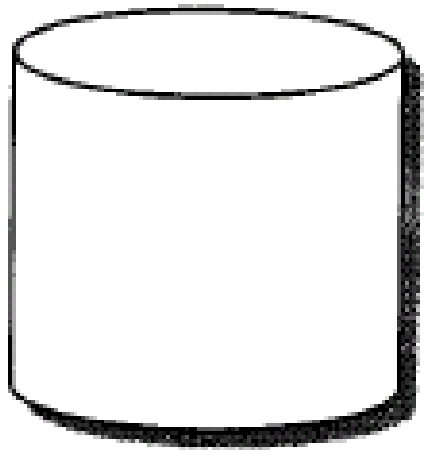
特点4.时变(cont.)

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

数据仓库支持管理决策

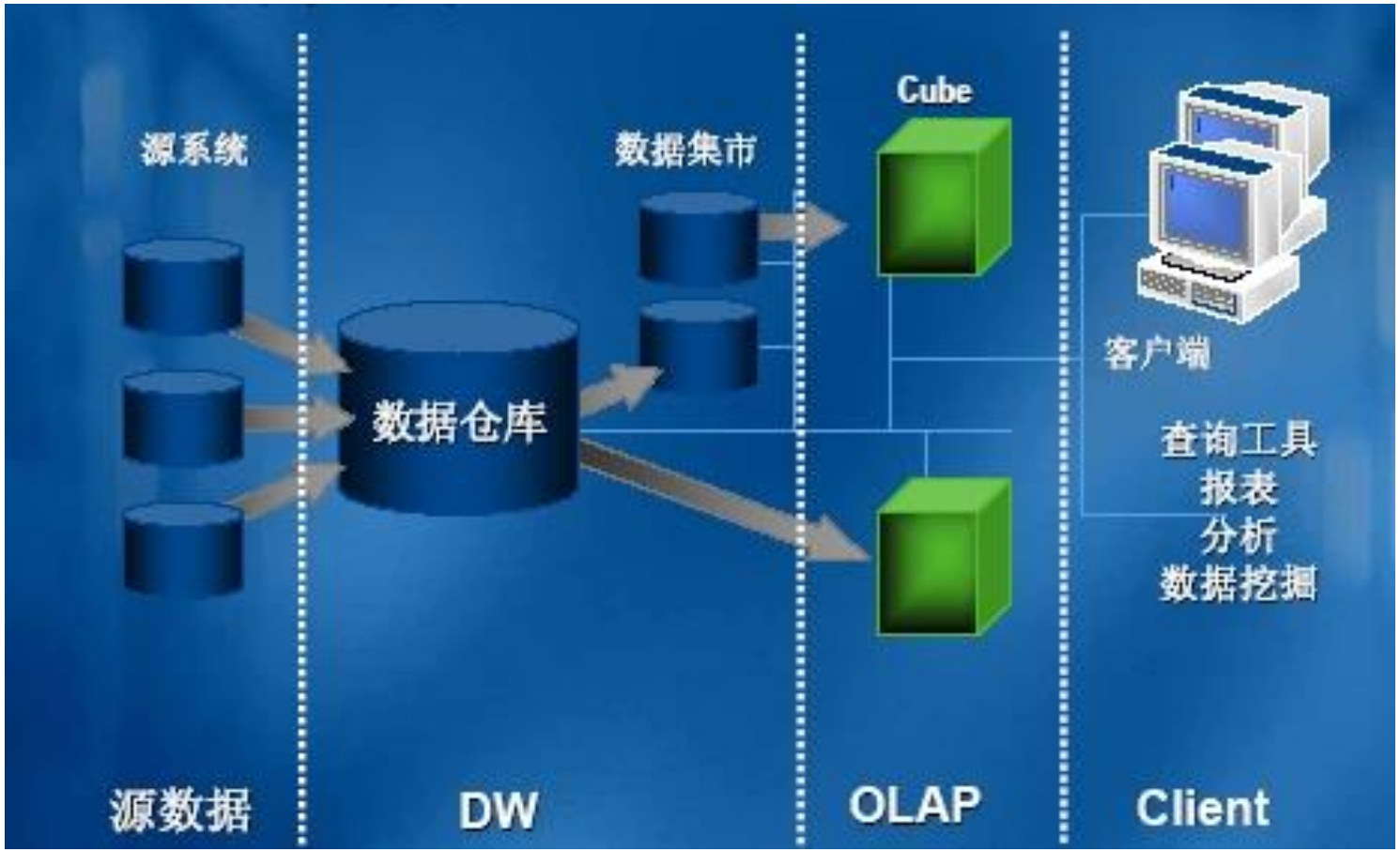
- 数据仓库支持OLAP、数据挖掘和决策分析
- OLAP从数据仓库中的综合数据出发，提供面向分析的多维模型，并使用多维分析的方法从多个角度、多个层次对多维数据进行分析，使决策者能够以更加自然的方式来分析数据
- 数据挖掘则以数据仓库和多维数据库中的数据为基础，发现数据中的潜在模式和进行预测
- 因此，数据仓库的功能是支持管理层进行科学决策，而不是事务处理

数据仓库的技术要求

厦門大學 XIAMEN UNIVERSITY

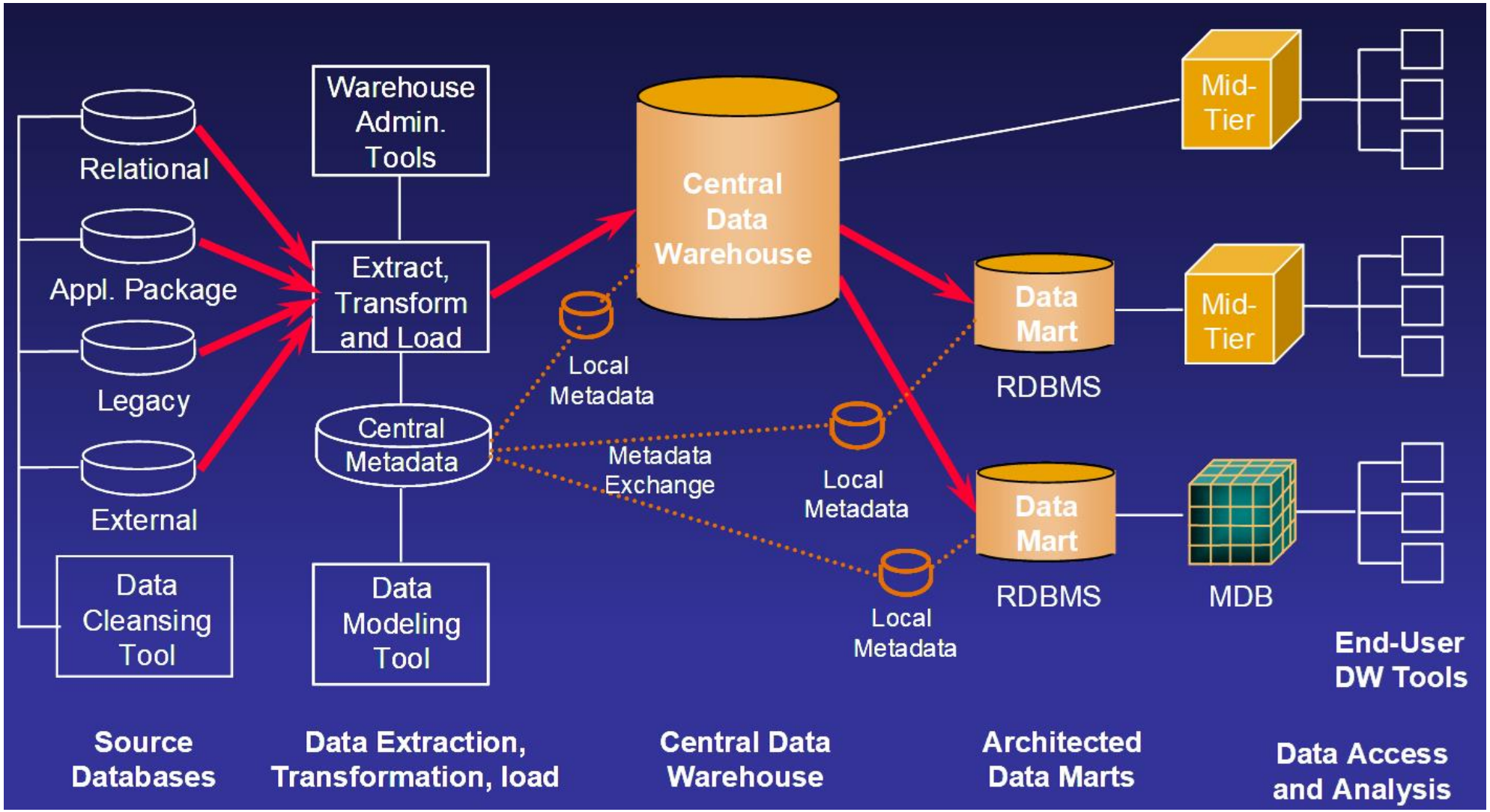
- 大量数据的组织和管理
  - 包含了大量的历史数据，它是从数据库中提取得来的，如何高效的组织和管理数据是数据仓库性能能否发挥的重要前提
- 复杂分析的高性能体现
  - 涉及大量数据的聚集、综合等，在进行复杂查询时经常会使用多表的联接、累计、分类、排序等操作
- 对提取出来的数据进行集成
  - 数据仓库中的数据是从多个应用领域中提取出来的，在不同的应用领域和不同的数据库系统中都有不同的结构和形式，所以如何对数据进行集成也是构建数据仓库的一个重要方面
- 对进行高层决策的最终用户的界面支持
  - 提供各种分析应用工具

数据仓库的体系结构

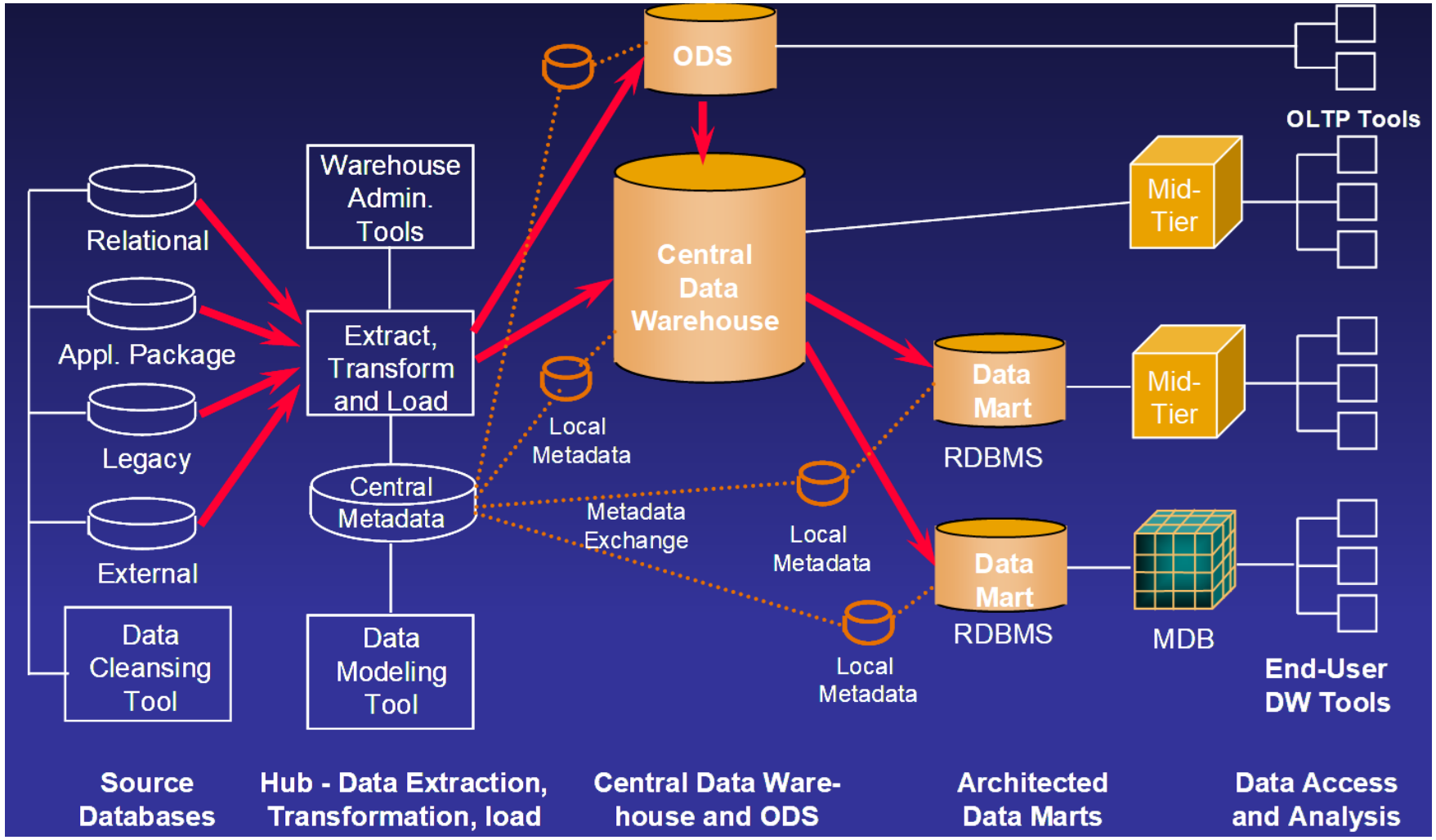


传统通用数据仓库体系结构

数据仓库的体系结构(cont.)

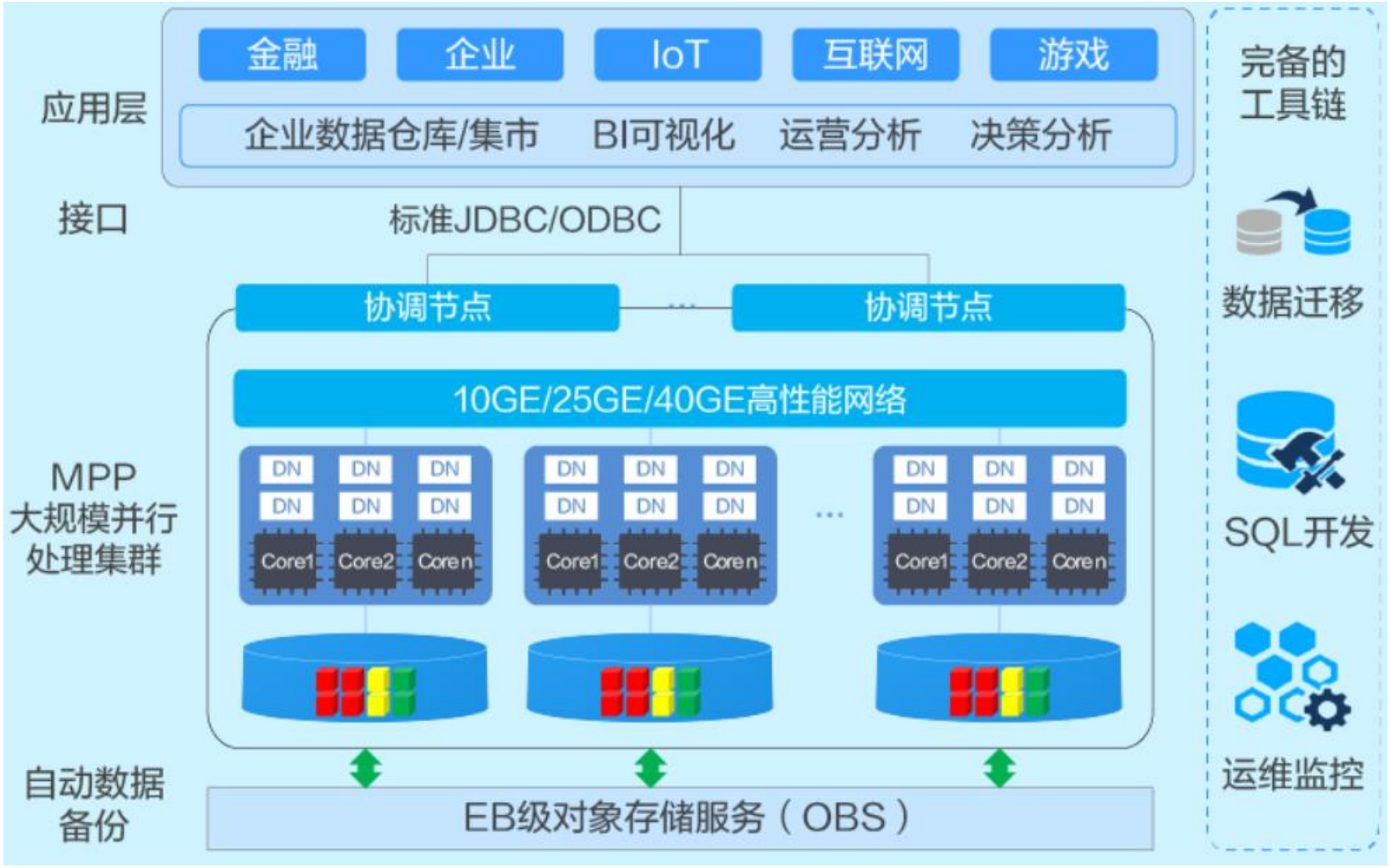


数据仓库的体系结构(cont.)



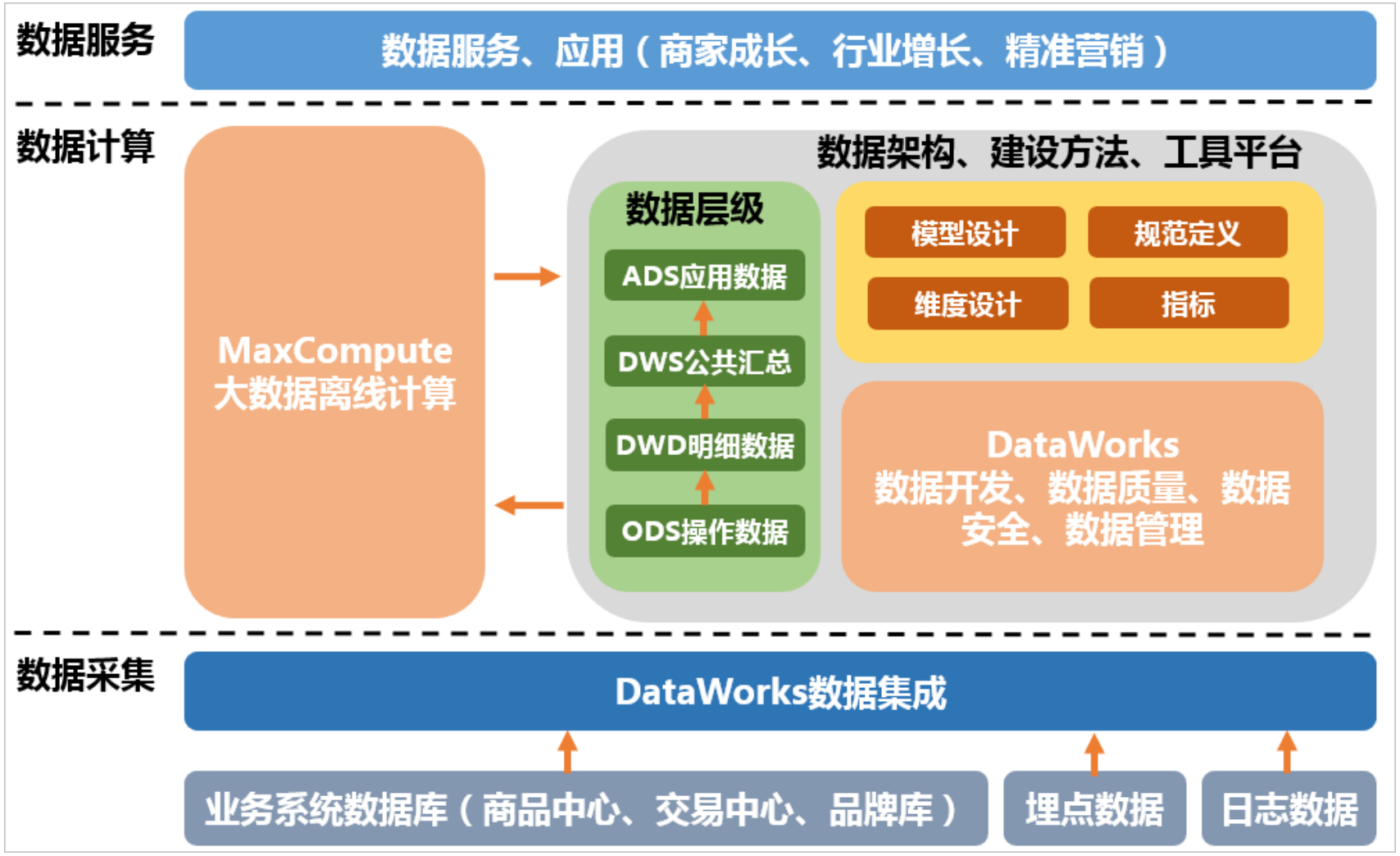
带ODS的数据仓库体系结构

华为云数仓GaussDB(DWS)的技术架构



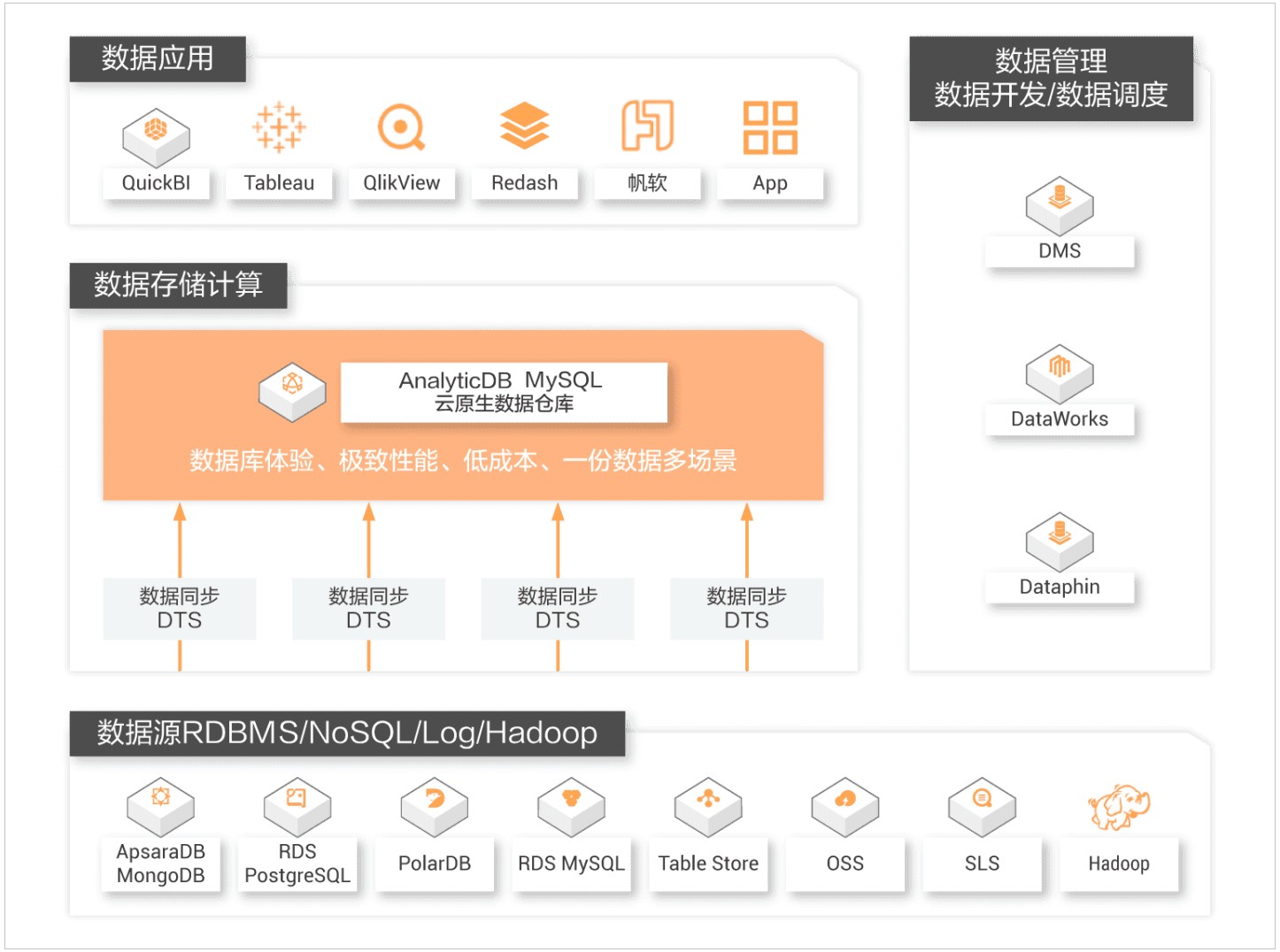
https://support.huaweicloud.com/intl/zh-cn/productdesc-dws/dws_01_1110.html

阿里云的离线数仓MaxCompute



https://blog.csdn.net/weixin_43114209/article/details/142081892

阿里云的实时数仓AnalyticDB



数据仓库体系结构之关键技术

- ETL
- 数据仓库存储
- 元数据
- 数据集市
- OLAP

ETL

- Extract(ion), Transform(ation), Load(ing)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

数据存储与元数据

- 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

数据集市与OLAP

厦門大學 XIAMEN UNIVERSITY

- 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

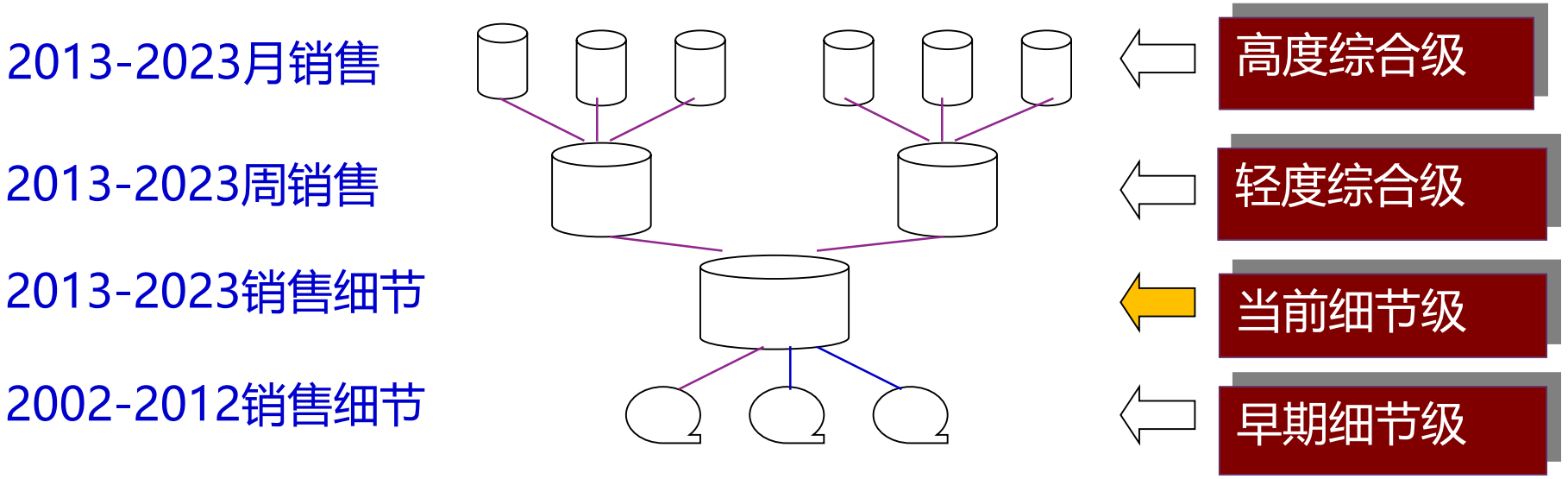
数据仓库的数据组织

厦門大學 XIAMEN UNIVERSITY

- 数据粒度
- 数据分割

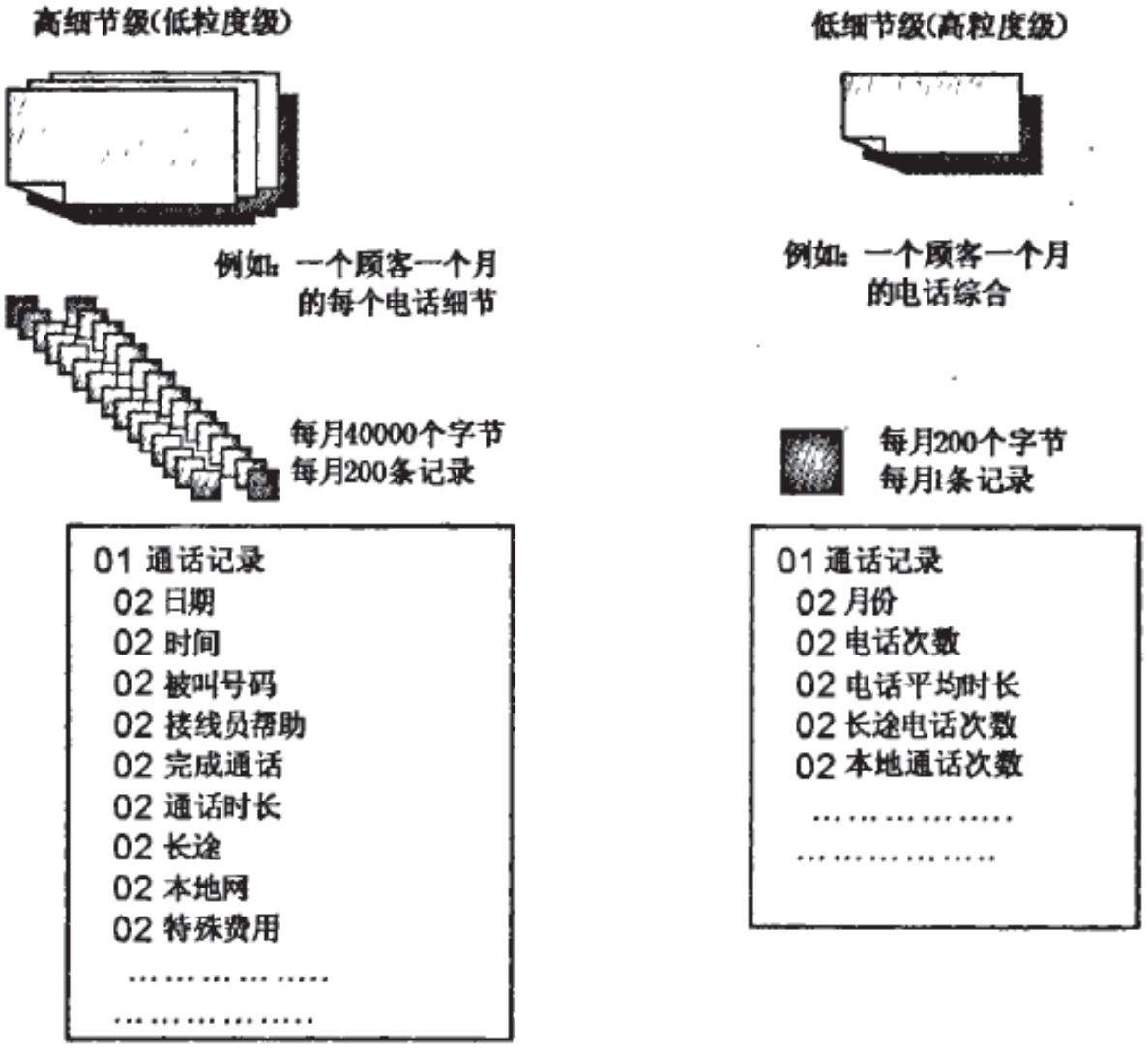
数据粒度

- 粒度(Granularity)是指数据仓库中数据的综合级别



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

粒度的两种不同表现形式

厦門大學 XIAMEN UNIVERSITY

- 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

数据仓库数据组织之数据分割

厦門大學 XIAMEN UNIVERSITY

- 数据分割(Partitioning)是数据仓库中数据存储的一个重要概念
  - 分割是将数据分散到各自的物理单元中去以便能分别处理，提高数据处理效率。
  - 数据分割后的数据单元称为分区。
  - 数据分割的目的
    - 便于进行数据的重构、索引、重组、恢复、监控、扫描

数据仓库数据组织之数据分割(cont.)

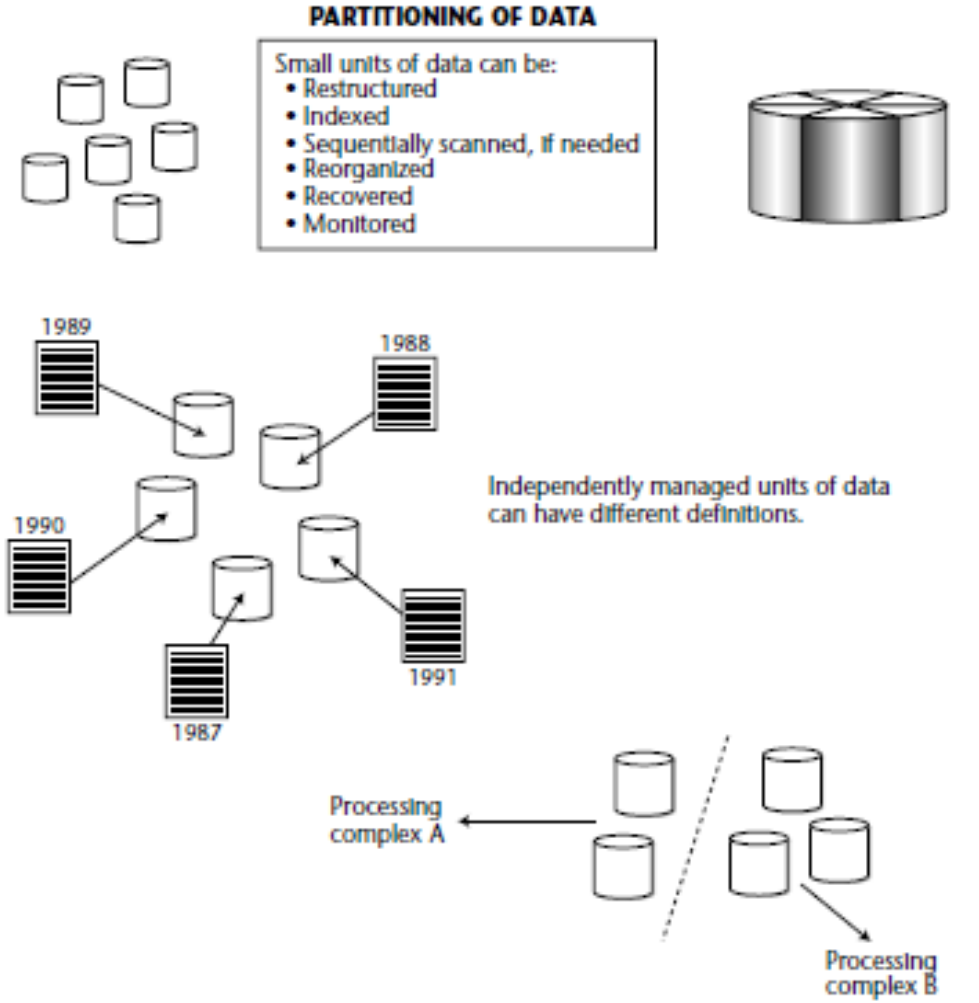


Figure 2-19 Independently managed partitions of data can be sent to different processing complexes with no other system considerations.

数据仓库数据组织之数据分割(cont.)

厦門大學 XIAMEN UNIVERSITY

- 数据分割的标准
  - 可按日期、地域、业务领域、组织单位或多个分割标准的组合
  - 一般分割标准种都包含日期
  - 示例：将人寿保险公司选择数据分割的物理单元
    - 2019年健康索赔
    - 2020年健康索赔
    - 2021年健康索赔
    - 2018年人寿保险索赔
    - 2019年人寿保险索赔
    - 2020年人寿保险索赔
    - 2021年人寿保险索赔
    - 2019年意外伤亡索赔
    - 2020年意外伤亡索赔
    - 2021年意外伤亡索赔

保险公司使用了日期即年和索赔类型作为标准来对数据分割

数据仓库数据组织之数据分割(cont.)

厦門大學 XIAMEN UNIVERSITY

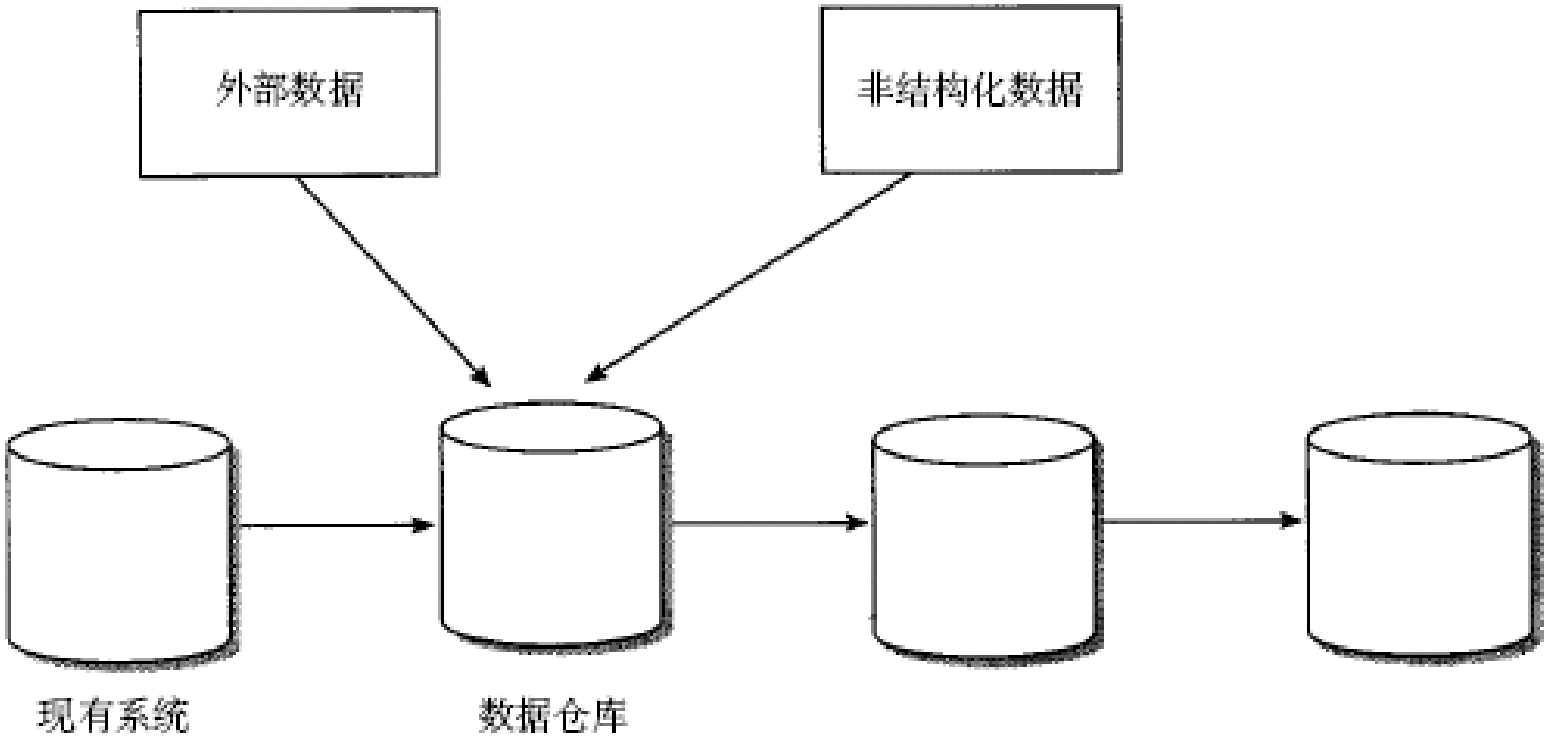
- 数据分割的方法
  - 垂直分割
    - 就是把一个表垂直分成两部分。这种类型的分割有助于把一大堆列分成两个独立的表，这两个表之间通过一个关键字段相关联
  - 水平分割
    - 就是把表按行分成两部分。这种类型的分割被用来存储与用户联系紧密的本地重要数据，从而减少网络查询
  - 图解分割
    - 经由多个分布系统把一个表分解成两部分。可以从指定的服务器或在多个服务器之间建立连接而得到一个表所需要的全部数据。这种类型的分割被用来把小的、静止的表从不稳定的、越变越大的表中分割出来

数据源

厦門大學 XIAMEN UNIVERSITY

- 数据源的种类
  - 内部数据源
    - 来自企业内部的数据，包括：业务系统中的结构化数据和非结构化数据
  - 外部数据源
    - 来自企业外部的数据，包括：外部的结构化数据和非结构化数据(居多)
    - 为确保获得正确的外部数据，需要建立可靠的监控方式和清洗
  - 结构化与非结构化数据
    - 结构化数据：关系数据库表示的数据
    - 非结构化数据：无法用关系数据库表示的数据，如图像、视频、声音和文字等
    - 半结构化数据：XML、JSON等文件

数据源(cont.)



ETL

厦門大學 XIAMEN UNIVERSITY

- ETL是构造数据仓库的前提，是数仓从数据源获得数据的必经之路，它包括四个过程：
  - 数据抽取
  - 数据转换
  - 数据清洗
  - 数据装载

ETL之数据抽取

厦門大學 XIAMEN UNIVERSITY

- 数据源确认
  - 确认数据的源系统和结构
- 抽取方法
  - 针对每个数据源，定义抽取过程是人工抽取还是基于工具的抽取（工具自己编写的还是购买的）
- 抽取频率
  - 对于每个数据源，确定数据抽取的频率，每天、每星期、每季度等
- 时间窗口
  - 对于每个数据源，表示出抽取过程进行的时间窗口，如T+1
- 工作顺序
  - 决定抽取任务中某项工作是否必须等到前面工作成功完成，才能开始
- 异常处理
  - 决定如何处理无法完成抽取的输入记录

ETL之数据抽取(cont.)

厦門大學 XIAMEN UNIVERSITY

- 从数据源中捕获的数据可分为：
  - 静态数据
    - 一般用于在数据仓库初始装载的时候进行，是相关数据源在某个时刻的快照
- 修正数据
  - 追加的数据捕获，是最后一次捕获数据后的修正
  - 追加的数据捕获可能是立刻进行的或者延缓进行的
    - 立即型数据捕获: 数据抽取发生在源系统中发生交易的时候，数据抽取是即时的或者实时的
    - 延缓型数据捕获：非即时的或实时的数据抽取

ETL之数据抽取(cont.)

厦門大學 XIAMEN UNIVERSITY

- 立即型数据抽取
  - 通过交易日志捕获数据：日志就是DBMS为应付突发情况的备份
    - 没有额外开销。需要保证日志刷新之前，已抽取了所有记录。
    - 缺点：如果源数据不是基于DB的则无法进行此方式的数据捕获
  - 从数据库触发器中捕获数据
    - 缺点1：只能捕获基于DB的数据
    - 缺点2：建立和维护触发器以及触发器的执行增加了开销
  - 从源应用程序中捕获数据
    - 优点：适用于所有的系统（基于DB的或者文件系统的）
    - 缺点： 1.程序的开销 2.可能会降低应用程序的性能

ETL之数据抽取(cont.)

厦門大學 XIAMEN UNIVERSITY

- 延缓型数据抽取
  - 基于日期和时间标记的捕获
    - 通过日期比较来选择应该抽取的数据：前提是源系统中有时间戳
    - 记录删除了如何抽取？删除先做标记(逻辑删除)，待抽取后物理删除，增加了开销
  - 通过文件的比较来捕获
    - 保存副本，然后比较昨天的副本和今天的副本以决定抽取那些数据。
    - 缺点：如果数据文件很大，则比较费时间
    - 优点：对于没有交易日志或者时间标记的而言，唯一可行的方法

ETL之数据转换与清洗

- 对从业务系统中抽取的数据根据数据仓库系统模型的要求，进行数据的转换、清洗、拆分、汇总等处理，保证来自不同系统、不同格式的数据具有一致性和完整性，并按要求装入数据仓库

- 主要完成以下原因引起的不一致问题
  - 源数据系统同数据仓库系统在模型上的差异
  - 数据源系统不一致，数据源可能基于不同平台数据库的数据，存在大量的转换工作
  - 源数据定义不规范导致错误数据
  - 对数据的约束不严格，导致无意义数据
  - 存在重复记录

ETL之数据转换与清洗(cont.)

- 数据转换和清洗工作可以在以下环节中实现
  - 在数据抽取过程中进行，要考虑抽取的性能和对业务系统性能的影响
  - 使用异步数据装载，以文件的方式处理。要考虑中间磁盘的存储量、ETL过程中的协调性、大量非SQL语句的编程
  - 在数据装载过程中进行数据处理
    - 要考虑装载的性能
  - 进入数据仓库以后再进行处理
    - 要考虑数据仓库引擎的海量数据处理性能

ETL之数据装载

- 数据装载是将从数据源中抽取、转换和清洗的数据装载到数据仓库系统中的过程

- ETL的作用：
  - 解决数据分散问题
    - ETL将多个业务系统中的数据集中起来，便于分析
  - 解决数据不清洁的问题
    - 分散的数据也带来了数据不清洁问题，如客户信息在不同系统中不一致
    - 分散的业务数据是面向业务的，而不是面向决策的，ETL可以进行转换
    - 转换后的数据设计成多维结构
  - 方便企业各个部门构建数据集市
    - 数据仓库是面向企业的应用
    - 针对各个部门的信息应用是构建数据集市
    - 数据集市是按照部门从数据仓库中抽取，并进行加工处理
    - 构建数据集市过程中，使用ETL，可以简化操作，提高效率

ETL之数据装载(cont.)

- 两种主要装载技术
  - 使用数据仓库引擎厂商提供的数据装载工具进行数据装载
  - 使用数据仓库引擎厂商提供的API编程进行数据装载

- ETL工具分类
  - 专业ETL厂商和产品，如Ascential Datastage
  - 整体解决方案提供商和产品，提供数据仓库存储、设计和展现工具的同时也提供ETL工具，但结构相对封闭，只支持自己的产品，如IBM InfoSphere DataStage, Oracle Warehouse Builder, Microsoft DTS等
  - Pentaho的Kettle（开源）

ETL之数据装载(cont.)

厦門大學 XIAMEN UNIVERSITY

- ETL工具选择
  - 对平台的支持（高性能的硬件和主机）
  - 对数据源的支持
  - 数据转换功能
  - 管理和调度功能
  - 集成和开放性
  - 对元数据的管理

ETL之数据装载(cont.)

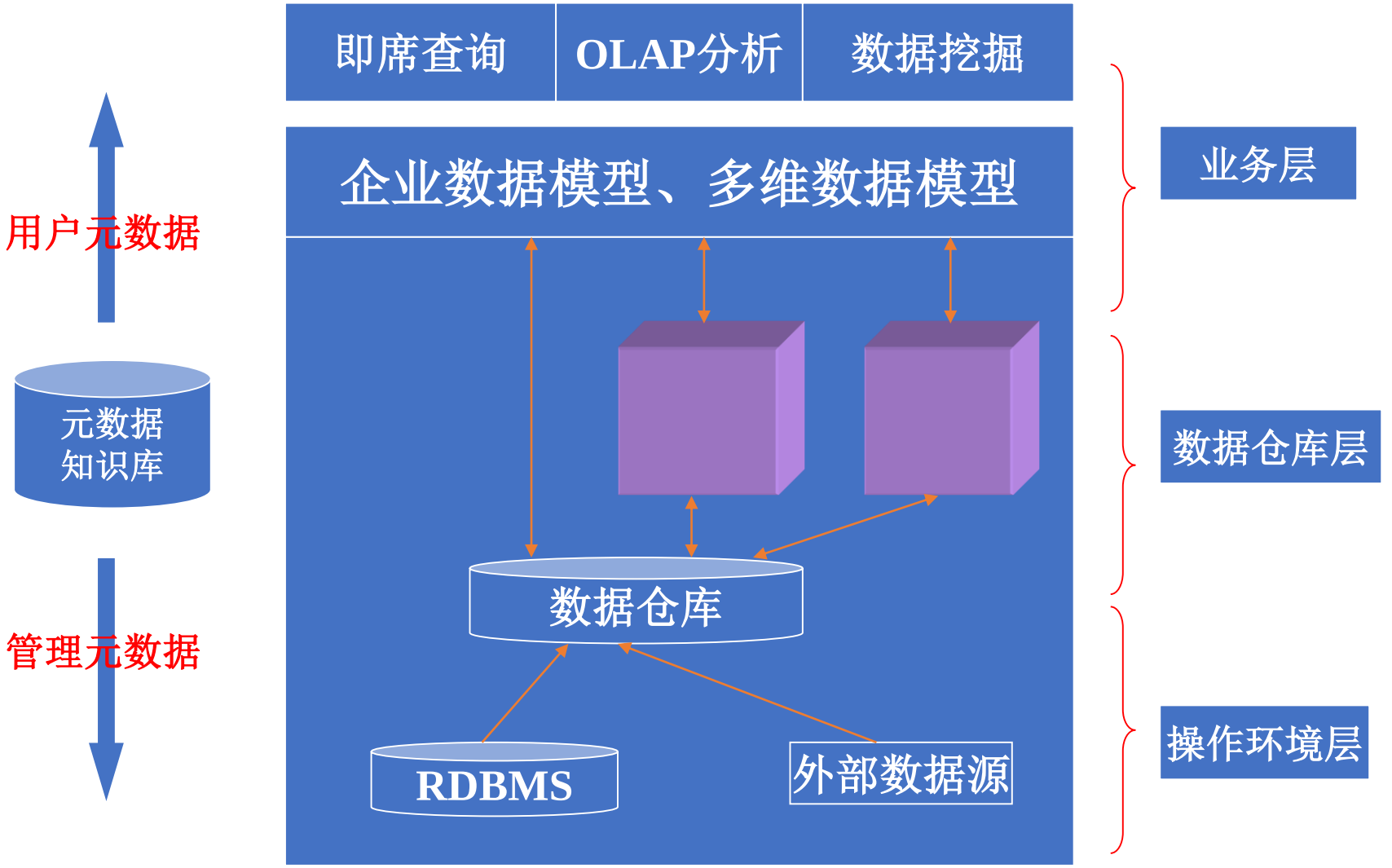
厦門大學 XIAMEN UNIVERSITY

- 装载策略需要考虑装载周期和数据追加策略
  - 装载周期要综合考虑业务分析需求和系统装载的代价，对不同业务系统的数据采用不同的装载周期
  - 数据追加策略根据抽取策略和业务规则确定，三种类型
    - 直接追加：直接将数据追加到目标表中
    - 全部覆盖：如果抽取数据包含了当前和所有历史状况，可以全部覆盖
    - 更新追加：连续记录业务的状态变化，并用当前最新状态同历史状态进行对比时，可以采用更新追加的方式

元数据(Meta data)

- 元数据是指描述数据的数据
  - 只要有程序和数据，元数据就会在信息处理环境中存在
- 在数仓中，元数据用来描述和定位其中的数据、数据来源及在数仓建设过程中的活动；还有数据的数据结构和相关操作的相关描述(输入、计算和输出)。
  - 数据从哪里来？更新频率多大？数据元素的含义是什么？进行了哪些计算、转换和筛选？
- 元数据管理贯穿于整个数据仓库构建和应用过程中
- 元数据可用文件存在元数据库中
- 要有效地管理数仓，必须设计一个描述能力强、内容完善的元数据

元数据(cont.)


即席查询
OLAP分析
数据挖掘
企业数据模型、多维数据模型
用户元数据
元数据
知识库
管理元数据
数据仓库
RDBMS
外部数据源
业务层
数据仓库层
操作环境层

元数据分类

- 按用途分
  - 管理元数据
    - 为负责开发、维护数仓IT人员所使用，用于开发和管理数仓
      - 数据仓库结构描述，包括模式、视图、层次结构等
      - 汇总算法，包括度量和维定义算法、数据粒度、主题域等
      - 操作环境到数据仓库的映射，包括ETL规则、安全策略等
  - 用户元数据
    - 从最终用户的角度来描述数仓
      - 如何连接数仓
      - 可用访问数仓的哪些部分

元数据分类(cont.)

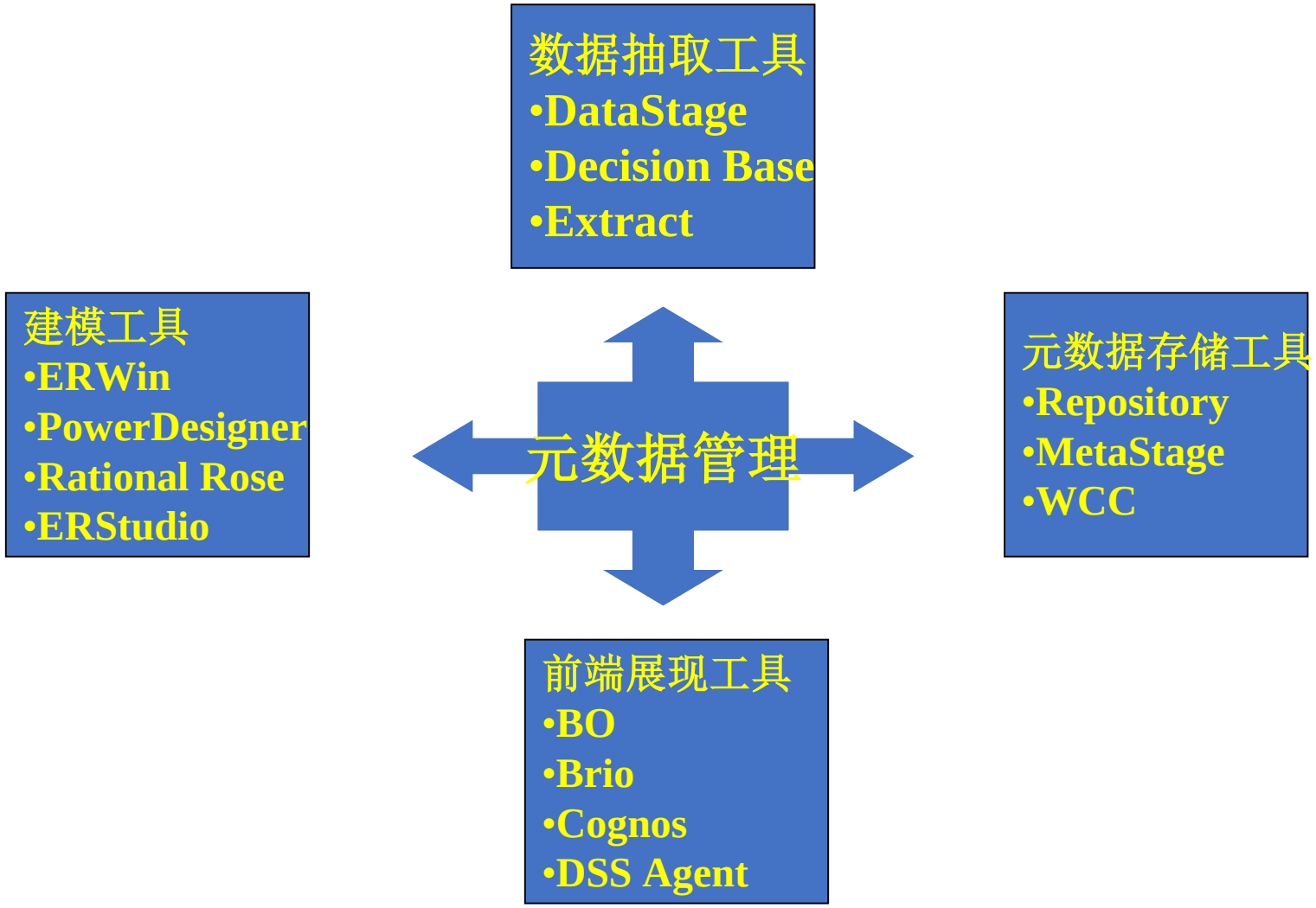
厦門大學 XIAMEN UNIVERSITY

- 其他分类方法
  - 按元数据来源
    - 数据源元数据、数据模型元数据、数据源与数仓映射元数据、数仓应用元数据
  - 按元数据的生成/使用时间
    - 设计时元数据、构建时元数据、运行时元数据
  - 按数仓功能区域
    - 数据获取元数据、数据存储元数据、信息传递元数据
  - 按元数据在数仓中承担的任务
    - 静态元数据、动态元数据

元数据的作用

- 元数据起到承上启下的作用
  - 元数据是进行数据集成所必需的
    - 数据源与数据仓库中数据的对应关系及转换规则都在元数据存储
  - 元数据的语义层可以帮助最终用户理解数据仓库中的数据
    - 元数据实现业务模型和数据模型之间的映射
  - 元数据是保证数据质量的关键
    - 减少使用者对数据结果的怀疑
  - 元数据可以支持需求变化
    - 元数据管理可以把整个业务的工作流、数据流和信息流管理起来

元数据的管理工具


数据抽取工具
•DataStage
•Decision Base
•Extract
建模工具
•ERWin
•PowerDesigner
•Rational Rose
•ERStudio
元数据管理
元数据存储工具
•Repository
•MetaStage
•WCC
前端展现工具
•BO
•Brio
•Cognos
•DSS Agent

数据存储

厦門大學 XIAMEN UNIVERSITY

- 数据存储的类型
  - 虚拟存储方式
    - 数据仓库中的数据仍然存储在源数据库中，只是根据用户的多维分析需求而形成多维视图，临时在源数据库中找出并提取所需要的数据，完成多维分析。
      - 优点：比较简单、花费少、使用灵活
      - 缺点：数据的净化、提取、集成需要花费大量的时间，在实际应用中这种方式难以建立起有效的、为决策服务的数据支持。
  - 基于关系表的存储方式
    - 基于关系表的存储方式是将数据仓库的数据存储在关系型数据库的表结构中，在元数据的管理下完成数据仓库的功能。
  - 多维数据库存储方式
    - 多维数据库的组织方式是直接面向OLAP分析操作的数据组织形式。这种数据库产品也比较多，实现方法也不尽相同。
    - 其数据组织采用多维数据结构文件进行存储，并有维索引及相应的元数据与其对应

数据文件的存储方式

- 简单堆积文件
  - 它将每天从数据库中提取加工后的数据逐日积累的存储起来
  - 按这种方式存储的数据细节化程度很高，可以应付多种细节查询，但分析时，查询的效率较低

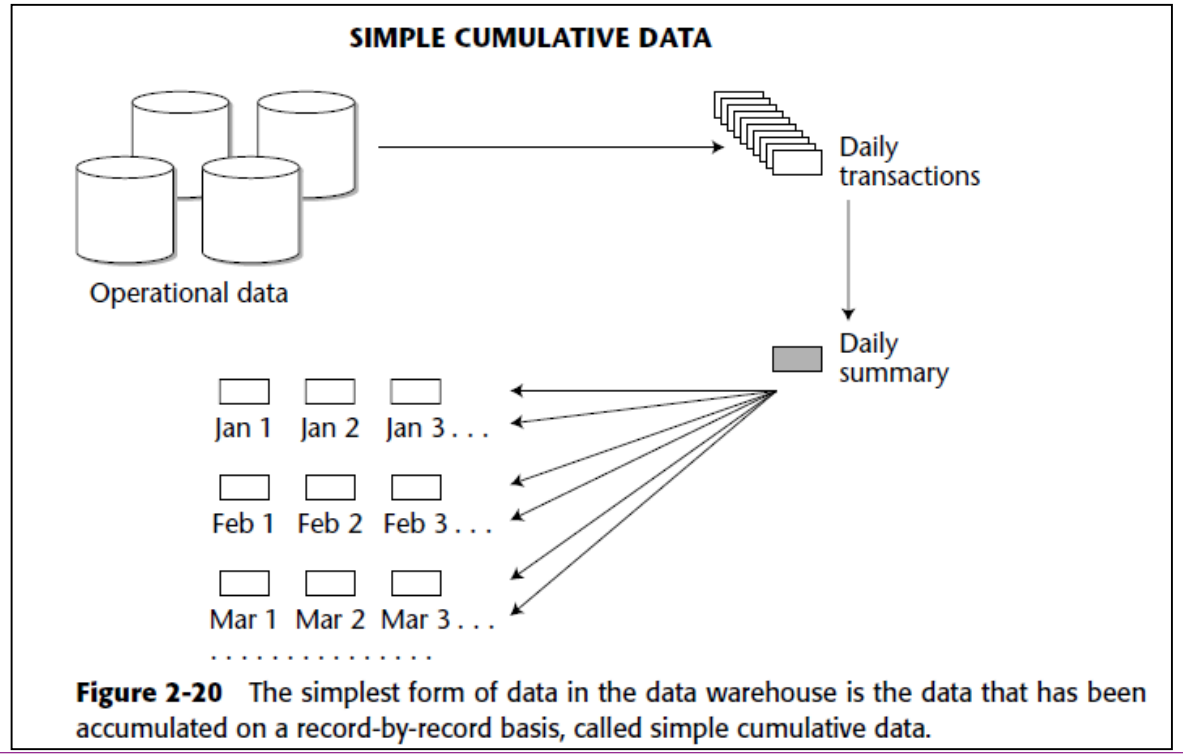


Figure 2-20 The simplest form of data in the data warehouse is the data that has been accumulated on a record-by-record basis, called simple cumulative data.

数据文件的存储方式(cont.)

- 轮转综合文件(Rolling summary data)
  - 它将数据按不同的期限轮转地存储
    - 例如，可将每一天的数据记录在一个日记录集中，当到达一个星期后再将这七天的数据进行综合然后存储在一个周记录中，同时将原来日记录集中的数据清空开始对新一周的每一天的数据进行记录；当到达一个月后，将周记录集中的数据进行综合然后存储在一个月记录中，而周记录中又开始新一个月的每一周的记录，以此类推
  - 按这种形式存储的数据较按简单堆积文件形式存储的数据其数据量大大减少，但是它是以损失细节程度为代价的，时间越久的数据，细节程度越低

数据文件的存储方式(cont.)

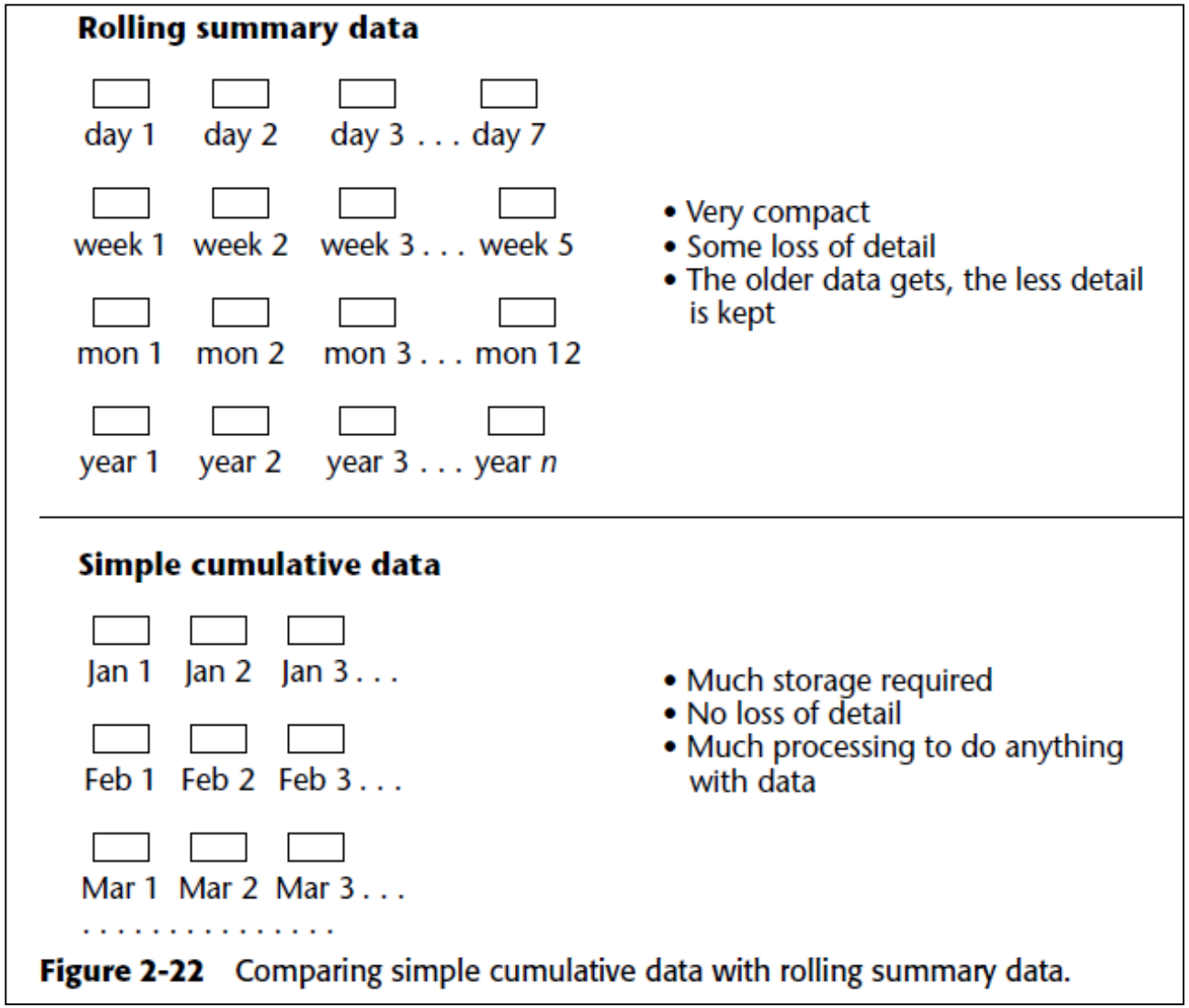
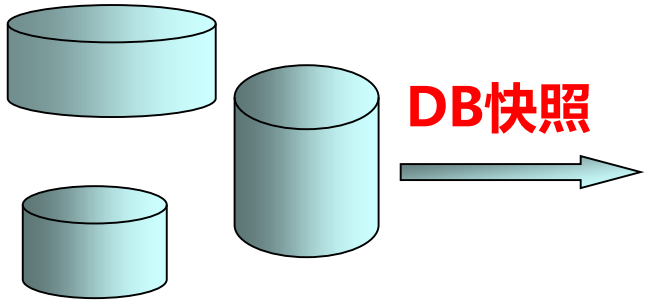


Figure 2-22 Comparing simple cumulative data with rolling summary data.

轮转综合文件与简单堆积文件比较

数据文件的存储方式(cont.)

- 简单直接文件
  - 数据是从操作型环境直接装入数据仓库中，并没有任何积累，只不过这种文件不是在每天的基础上组织的，而是以较长时间（如一个星期、一个月）为单位的
  - 因此，简单直接文件是按一定时间操作型数据库的一个快照，即按一定所时间间隔对数据库的采样



一月份顾客表

姓名	顾客号	地址
张平	C980100	北京
王英	C980101	天津
王宾	C980102	上海
刘仲	C980104	重庆
…	…	…

数据文件的存储方式(cont.)

- 连续文件
  - 它是通过比较两个连续的简单直接文件的不同而生成的另一种连续文件，生成的连续文件又可以和新的简单直接文件一起生成新的连续文件
    - 例如：通过比较两个简单文件“1月份顾客表”和“2月份顾客表”生成一个连续文件“1-2月份顾客表”，然后再通过比较连续文件“1-2月份顾客表”和另一个简单直接文件“3月份顾客表”生成一个相等连续文件“1-3月份顾客表”等

1月份顾客表

姓名	顾客号	地址
张平	C980100	北京
王英	C980101	天津
王宾	C980102	上海
刘仲	C980104	重庆

2月份顾客表

姓名	顾客号	地址
张平	C980100	北京
王英	C980101	沈阳
王宾	C980102	上海
刘仲	C980104	大连

1-2月份顾客表

姓名	顾客号	时间	地址
张平	C980100	1-2月	北京
王英	C980101	1-1月	天津
王英	C980101	2-2月	沈阳
王宾	C980102	1-2月	上海
刘仲	C980104	1-1月	重庆
刘仲	C980104	2-2月	大连

本章小结

- **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**
  - 数仓发展
  - 数仓的定义及4个特点
  - 数仓体系结构
  - 数据源
  - ETL
  - 数据存储

本章练习

1. 简述商务智能的核心要素，商务智能与数据仓库的关系是什么？
2. 为什么传统数据库称为操作型或事务的？为什么它们不适合数据分析？
3. 简述数据仓库的4个特点。
4. 联机分析处理(OLAP)的目的是什么？它与数据仓库的关系是什么？
5. 简述ETL过程。
6. 什么是元数据？它的作用是什么？
7. 哪些语言可用于查询数据仓库？（可能不止一种）
8. 什么是大数据(big data)？它与数据仓库的关系是什么？请给出例子。