数 据 仓 库

Data Warehouse

杨育彬

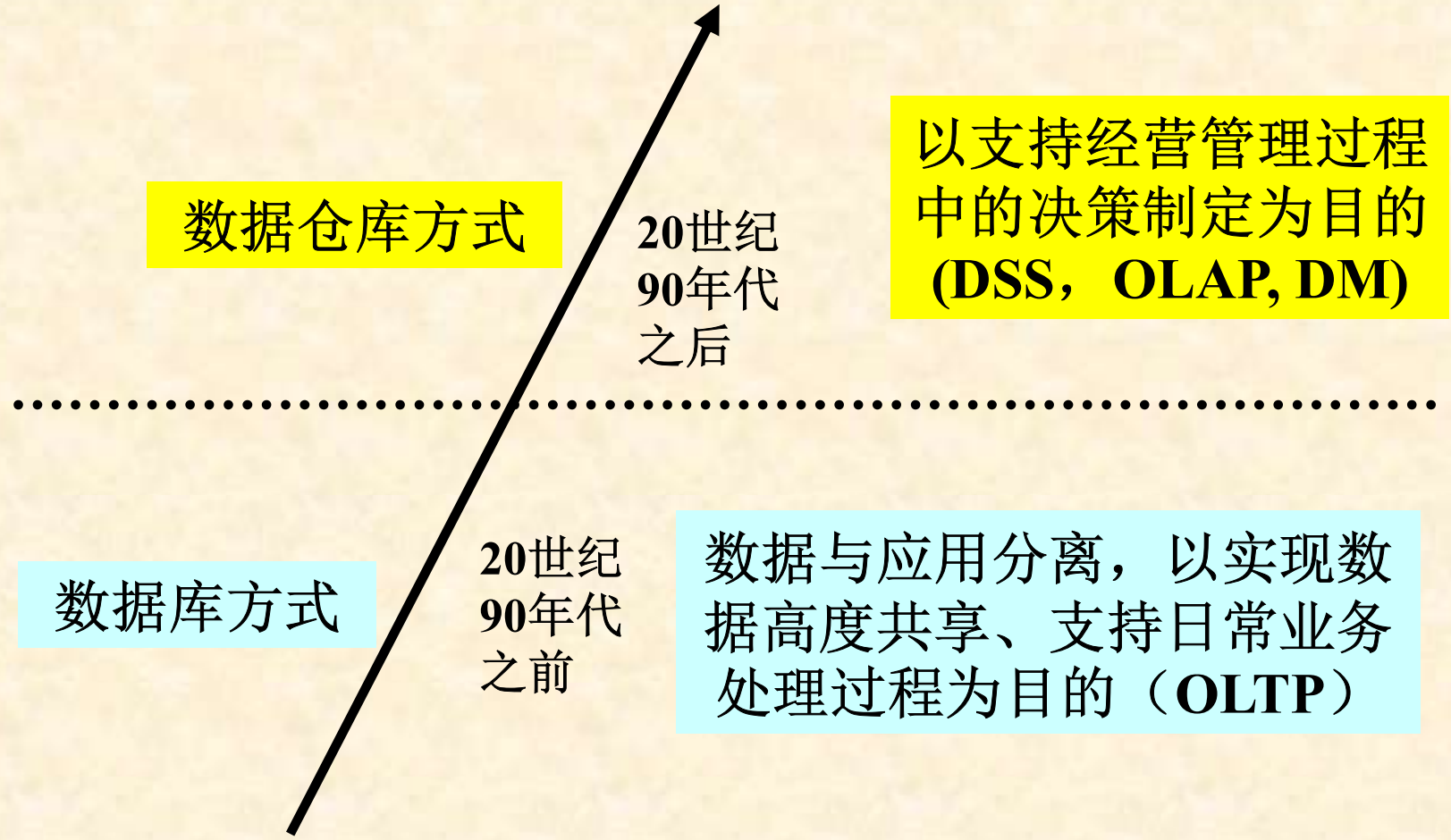
yangyubin@nju.edu.cn

数据仓库

1. 引　言　√
2. 从数据库到数据仓库
3. 数据分析与数据仓库
4. 数据仓库的四大特色
5. 数据仓库的基本结构
6. 数据仓库的设计
7. 联机分析处理 **(OLAP)**
8. 多维建模 **(Dimensional Modeling)**
9. 数据仓库的应用

1. 引 言

- 数据仓库的起因


数据仓库方式
20世纪
90年代
之后
以支持经营管理过程
中的决策制定为目的
(DSS，OLAP, DM)
数据库方式
20世纪
90年代
之前
数据与应用分离，以实现数
据高度共享、支持日常业务
处理过程为目的（OLTP）

1. 引 言

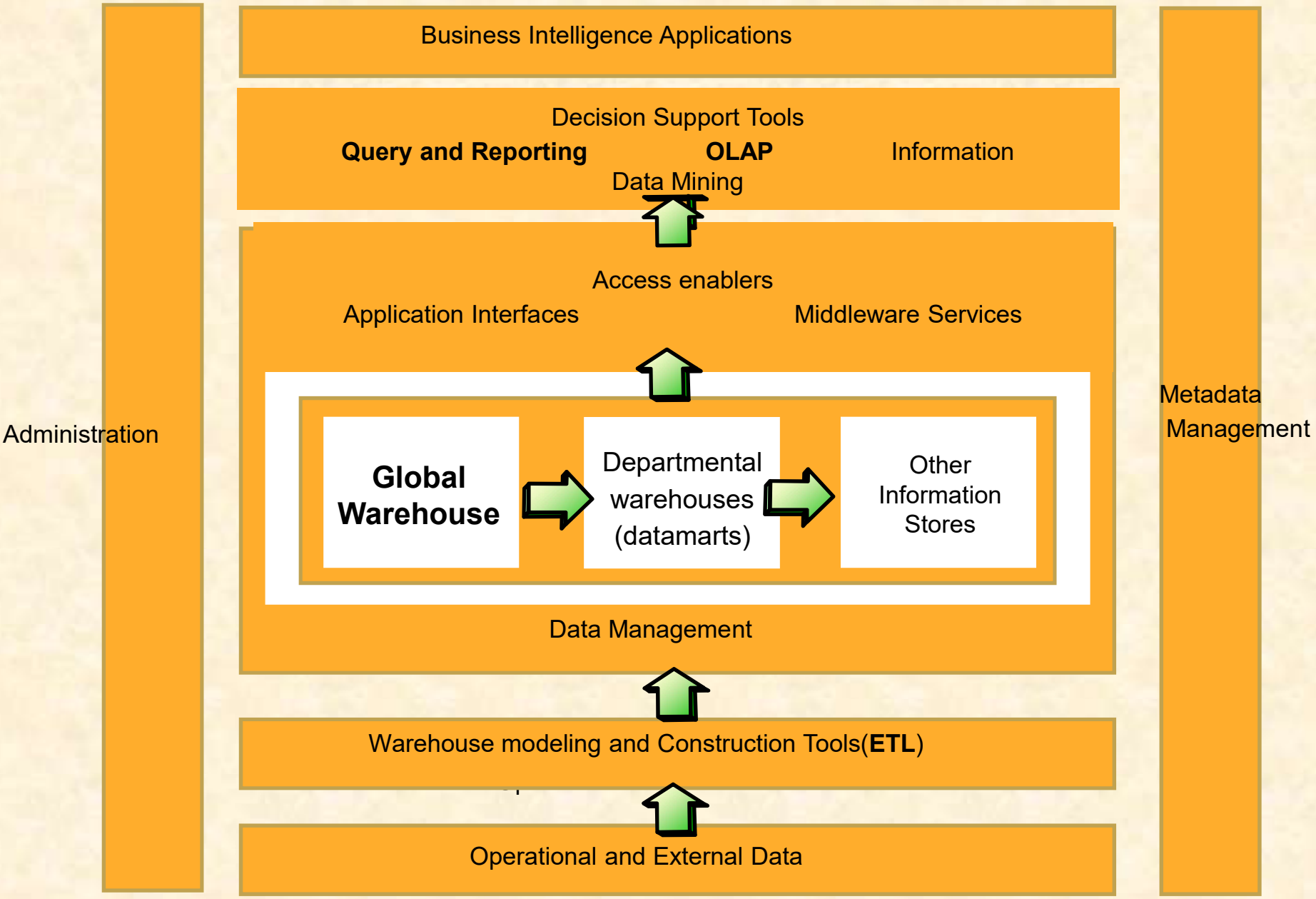
- 商业智能（BI）系统的需求
  - 企业的生存发展的关键在于：对不同的用户需求做出快速的反应及正确的决策，并提供优质的产品和服务。而这些反应都必须建立在对全面、准确和及时的信息基础上。（随需应变）
  - 存储数据的爆炸性增长激发了对新技术和自动工具的需求，以便将海量数据转换成信息和知识 （大数据）
  - 智能的表现
    - 信息共享和企业信息集成
    - 知识挖掘与管理
  - 商务智能不是通常的业务处理。它的目标是如何更快、更容易地做更好的决策。

1. 引 言

- 瓶颈问题：如何从数据到知识？
  - 数据仓库技术应运而生，考虑两个方面：
    - 如何有效地组织大量的数据，维护数据的一致性，方便用户的访问
    - 如何为决策人员有效地使用信息提供方便，通过使用数据仓库系统对企业的经营管理做出正确的决策，为企业带来经济效益。
      - OLAP + Data Mining

1. 引 言

- 商业智能（BI）系统的需求
  - 组成部分：
    - 数据仓库
    - 数据抽取/转换/加载(ETL)
    - 联机分析处理 (OLAP)
      - 广义：相对OLTP而言，包括查询、报表分析、OLAP分析和数据挖掘
      - 狭义：多维数据分析
    - 数据挖掘
    - 指标展现与分析(报表分析、可视化、多视图等)


Business Intelligence Applications
Decision Support Tools
Query and Reporting
OLAP
Information
Data Mining
Access enablers
Application Interfaces
Middleware Services
Administration
Global Warehouse
Departmental warehouses (datamarts)
Other Information Stores
Metadata Management
Data Management
Warehouse modeling and Construction Tools(ETL)
Operational and External Data

1. 引 言

- 商业智能（BI）系统的需求
  - 实施步骤：
    - 需求分析：明确的定义企业对商业智能的期望和需求，包括需要分析的主题，各主题可能查看的维度，即需要发现企业哪些方面的规律
    - 数据仓库建模：建立企业数据仓库的逻辑模型和物理模型，并规划系统的应用架构，将企业各类数据按照分析主题进行组织和归类

1. 引 言

- 商业智能（BI）系统的需求
  - 实施步骤：
    - 数据抽取：将数据从业务系统中进行抽取、净化、转换和装载，形成可以被系统识别的统一数据格式，导人数据仓库存放
    - 建立分析报表：展现和挖掘，生成报表，辅助做出下一步的生产营销决策
    - 数据测试与系统改进

1. 引 言

- 数据仓库的起因
  - “数据太多，信息不足”的现状(数据库系统的局限性)
    - 数据库适于存储高度结构化的日常事务细节数据，而决策型数据多为历史性、汇总性或计算性的多维性数据，多表现为静态数据，不需直接更新，但可周期性刷新。
    - 在事务处理环境中，决策者可能并不关心具体的细节信息，在决策分析环境中，如果这些细节数据量太大一方面会严重影响分析效率，另一方面这些细节数据会分散决策者的注意力。

1. 引 言

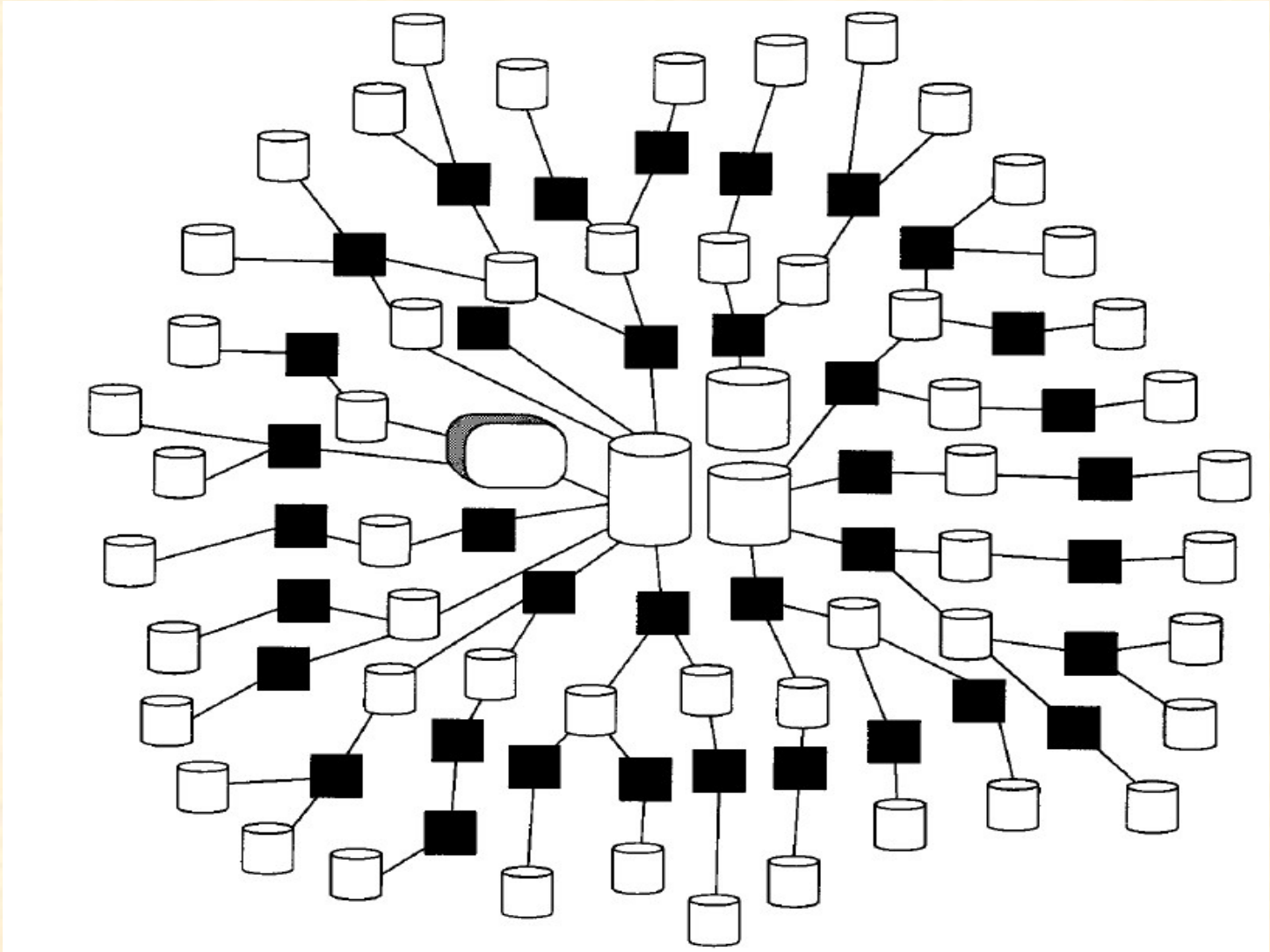
- 当事务型处理环境和分析型处理环境在同一个数据库系统中，事务型处理对数据的存取操作频率高，操作处理的时间短，而分析型处理可能需要连续运行几个小时，从而消耗大量的系统资源。

– 分散、异构环境的复杂数据源

- 决策型分析数据的数据量大，这些数据有来自企业内部的，也有来自企业外部的。来自企业外部的数据又可能来自不同的数据库系统，在分析时如果直接对这些数据操作会造成分析的混乱。
- 对于外部数据中的一些非结构化数据，数据库系统常常是无能为力。

1. 引 言

- “蜘蛛网”问题



1. 引 言

➢ “抽取”处理

- 需要总体分析数据时与联机事务处理性能不发生冲突
- 数据的控制方式发生了转变，最终用户 “拥有” 这些数据

问题：

- 数据可信性
- 生产率
- 数据转化为信息的不可行性

原因：

- 数据无时基
- 数据算法上的差异
- 抽取的多层次
- 外部数据问题
- 无起始公共数据源

1. 引 言

- 原始数据 **VS** 导出数据
  - 操作运行所用细节性数据 **VS** 统计计算出满足管理需要的数据

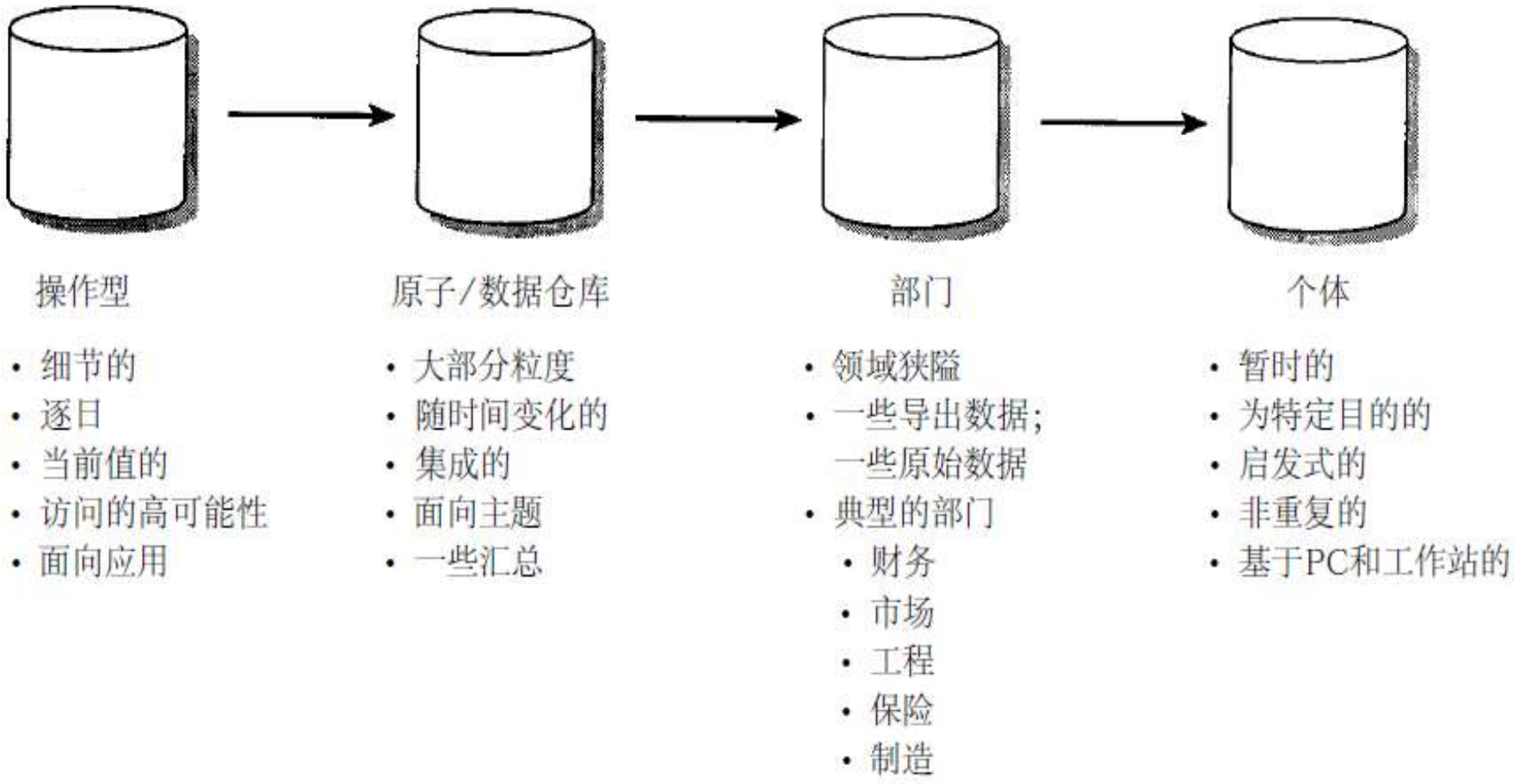
原始数据/操作型数据	导出数据/DSS数据
• 面向应用	• 面向主题
• 详细的	• 综合的,或提炼的
• 在存取瞬间是准确的	• 代表过去的数据
• 为日常工作服务	• 为管理者服务
• 可更新	• 不更新
• 重复运行	• 启发式运行
• 处理需求事先可知	• 处理需求事先不知道
• 生命周期符合 SDLC	• 完全不同的生命周期
• 对性能要求高	• 对性能要求宽松
• 一个时刻存取一个单元	• 一个时刻存取一个集合
• 事务处理驱动	• 分析处理驱动
• 更新控制主要涉及所有权	• 无更新控制问题
• 高可用性	• 松弛的可用性
• 整体管理	• 以子集管理
• 非冗余性	• 时常有冗余
• 静态结构；可变的内容	• 结构灵活
• 一次处理数据量小	• 一次处理数据量大
• 支持日常操作	• 支持管理需求
• 访问的高可能性	• 访问的低可能性或适度可能性

1. 引 言

- 事务处理环境不适宜DSS应用
  - ➢事务处理和分析处理的性能特性不同
    - 前者：适用于服务于单一目的的商务过程
    - 后者：从不同聚集层抽取数据，并使用高级方法分析企业数据
  - ➢数据的动态集成问题
  - ➢数据的综合问题（不需要太多细节数据）
  - ➢历史数据问题

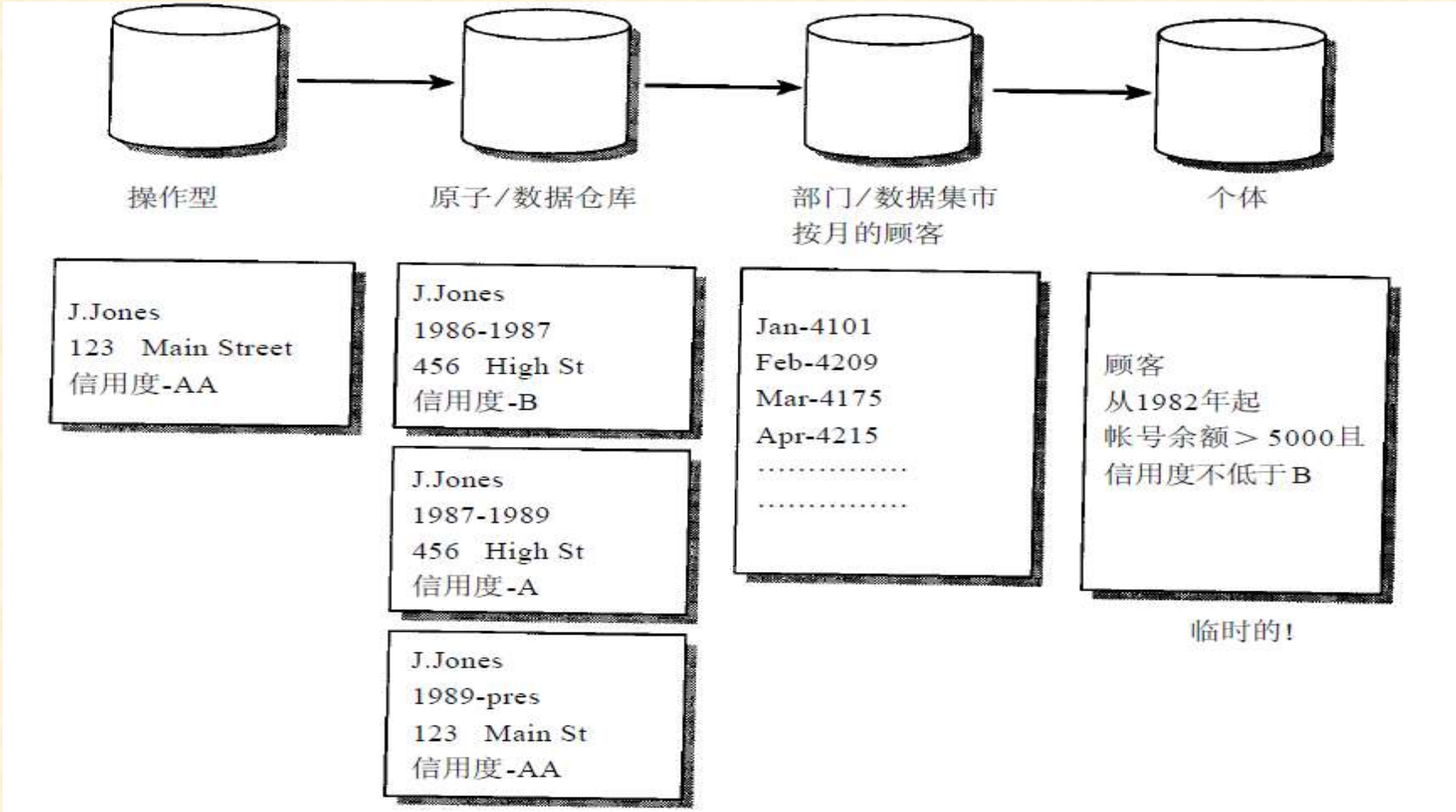
1. 引 言

❑数据分离的数据体系结构扩展



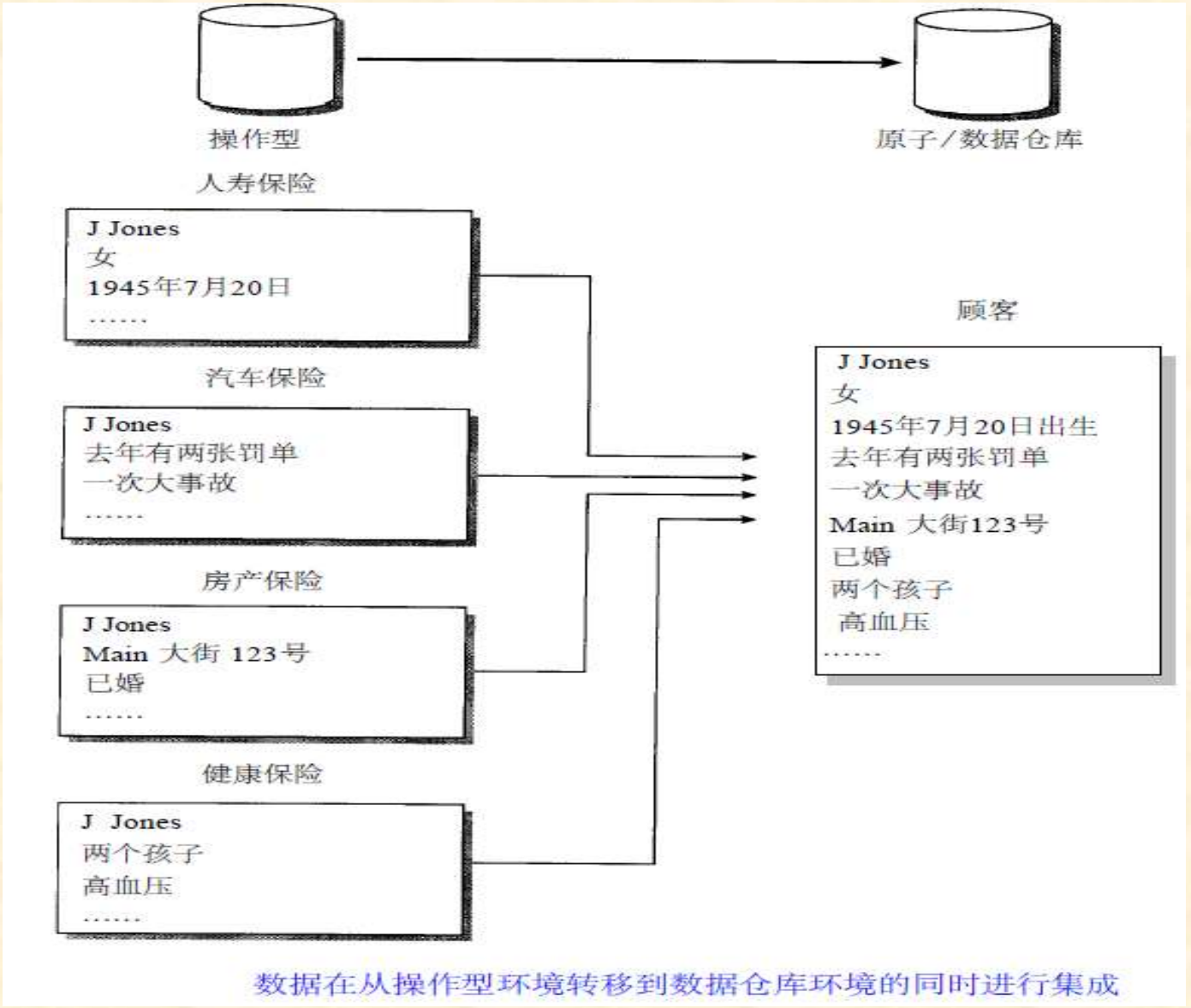
1. 引 言

- 例子：顾客分析


操作型
原子/数据仓库
部门/数据集市
按月的顾客
个体
J.Jones
123 Main Street
信用度-AA
J.Jones
1986-1987
456 High St
信用度-B
J.Jones
1987-1989
456 High St
信用度-A
J.Jones
1989-pres
123 Main St
信用度-AA
Jan-4101
Feb-4209
Mar-4175
Apr-4215
顾客
从1982年起
帐号余额＞5000且
信用度不低于B
临时的!
J.Jones现在的
信用度是多少？
J.Jones的信用
历史如何？
我们吸引越来越多
或越来越少的客户？
我们所分析的客户
趋势如何？
可用不同数据层次进行查询的不同类型

1. 引 言

- 顾客数据的集成



数据在从操作型环境转移到数据仓库环境的同时进行集成

1. 引 言

- 如何创建和使用数据仓库？
  - 如何获取数据？ →数据的抽取与刷新
  - 如何管理数据？ →数据的存储与管理
  - 数据的访问模式？ →数据的分析与挖掘
  - 结果的浏览方式？ →数据的展现

1. 引 言

- 在数据仓库中可以执行的数据分析操作
  - **query**
    - **SQL**查询和简单的统计查询
  - **reporting**
    - 执行预先定义好的查询命令，并以报表的格式返回查询结果
  - **statistical analysis**
    - 统计分析软件包：**SAS**，**SPSS**
  - **OLAP**
    - 多维统计分析
  - **data mining**
    - 数据挖掘

1. 引 言

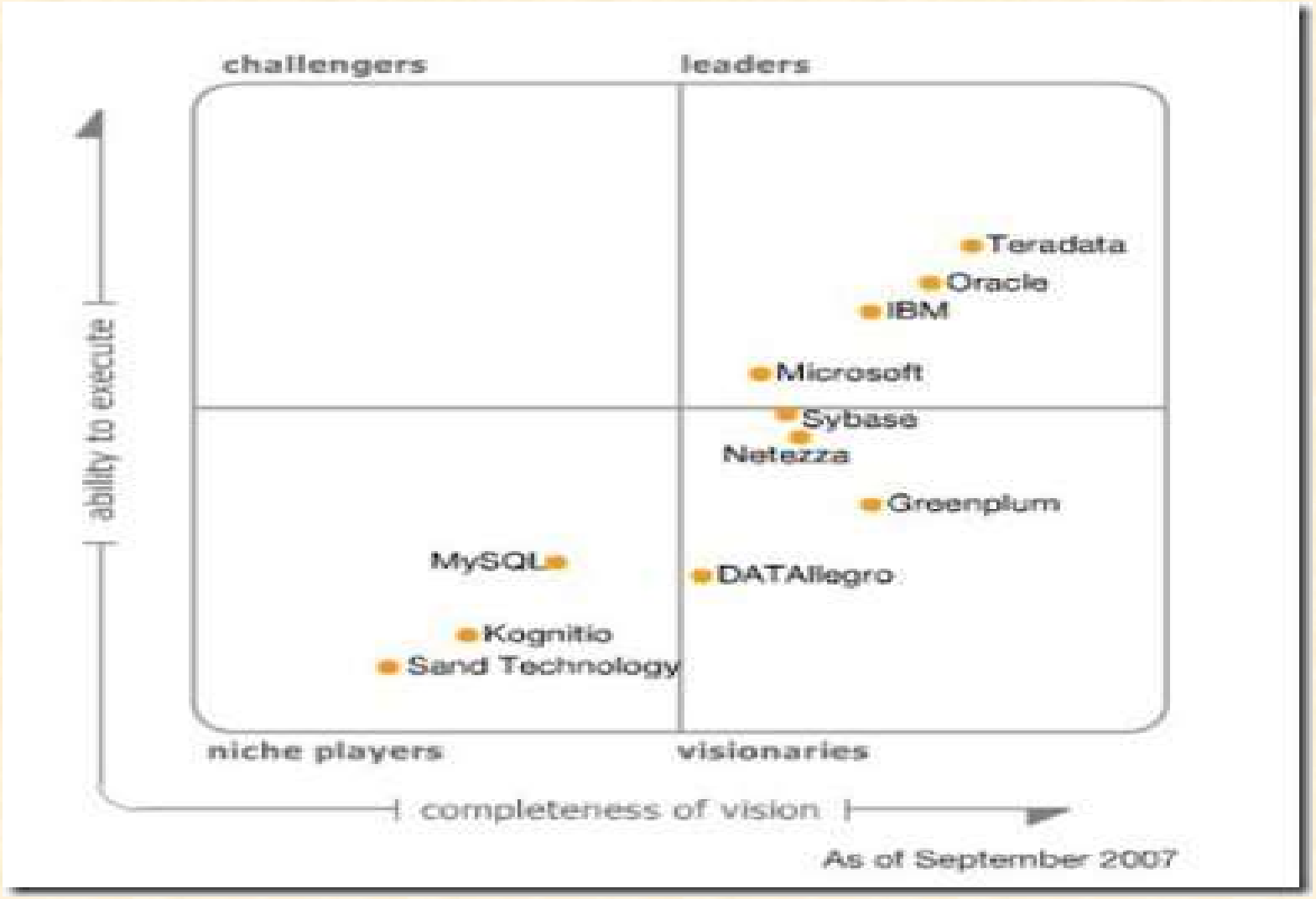
- 数据仓库的应用
  - 商品营销
  - 生产控制
  - 金融、电讯、保险等领域的业务策划、风险控制等职能应用

  ……

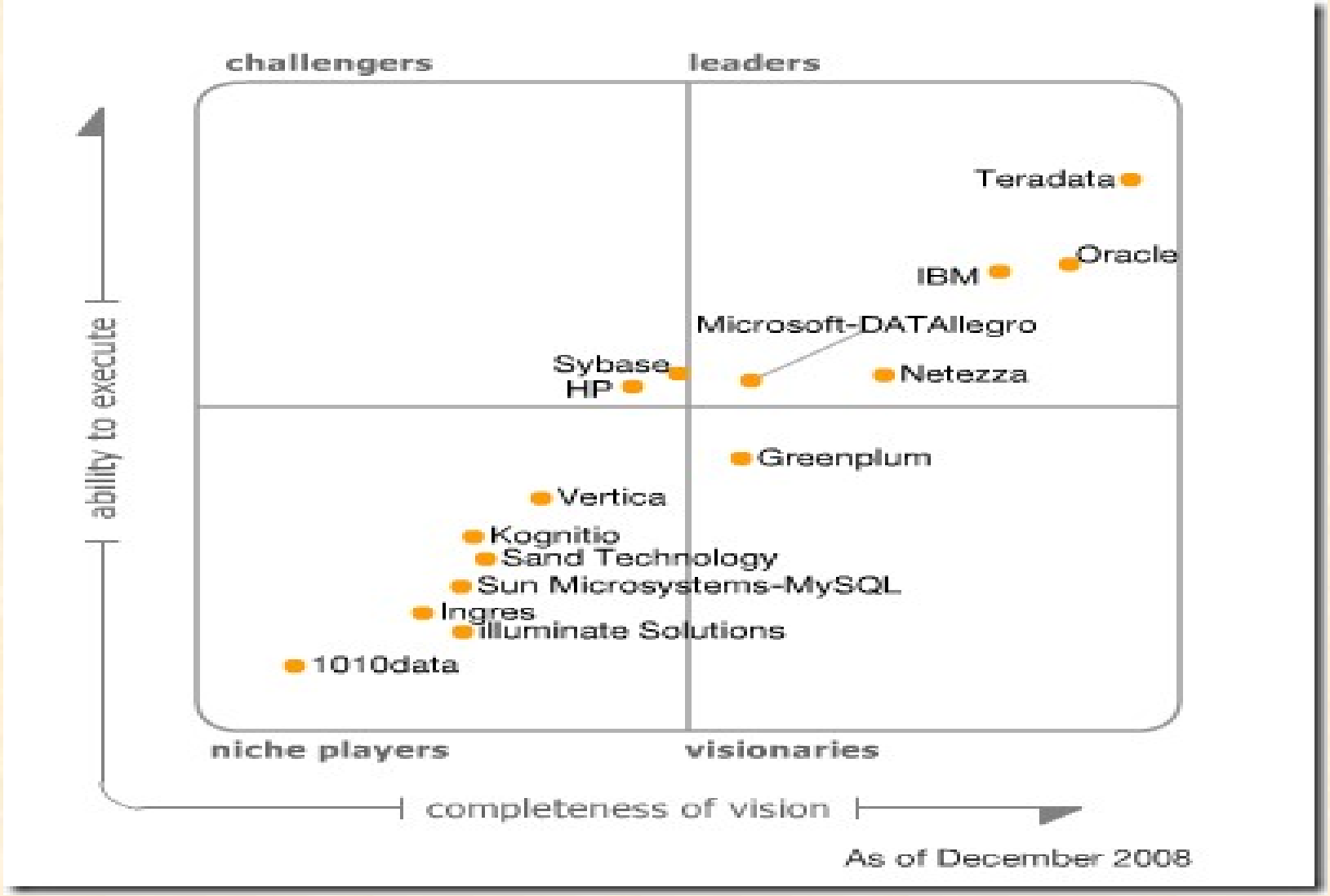
1. 引 言

- 数据仓库管理工具
  - Oracle：Enterprise Manager
  - Sybase：Warehouse Studio
  - IBM：DB2 Warehouse Manager
  - Microsoft: SQL Server DW/BI Toolkit
  - SAS：Warehouse Administrator
  - CA：PLATINUM ERWin、InfoPump
  - NCR：Database Manager
  - Teradata： EDW/ADW
  - SAP：Business Objects
  - AWS Redshift, ……

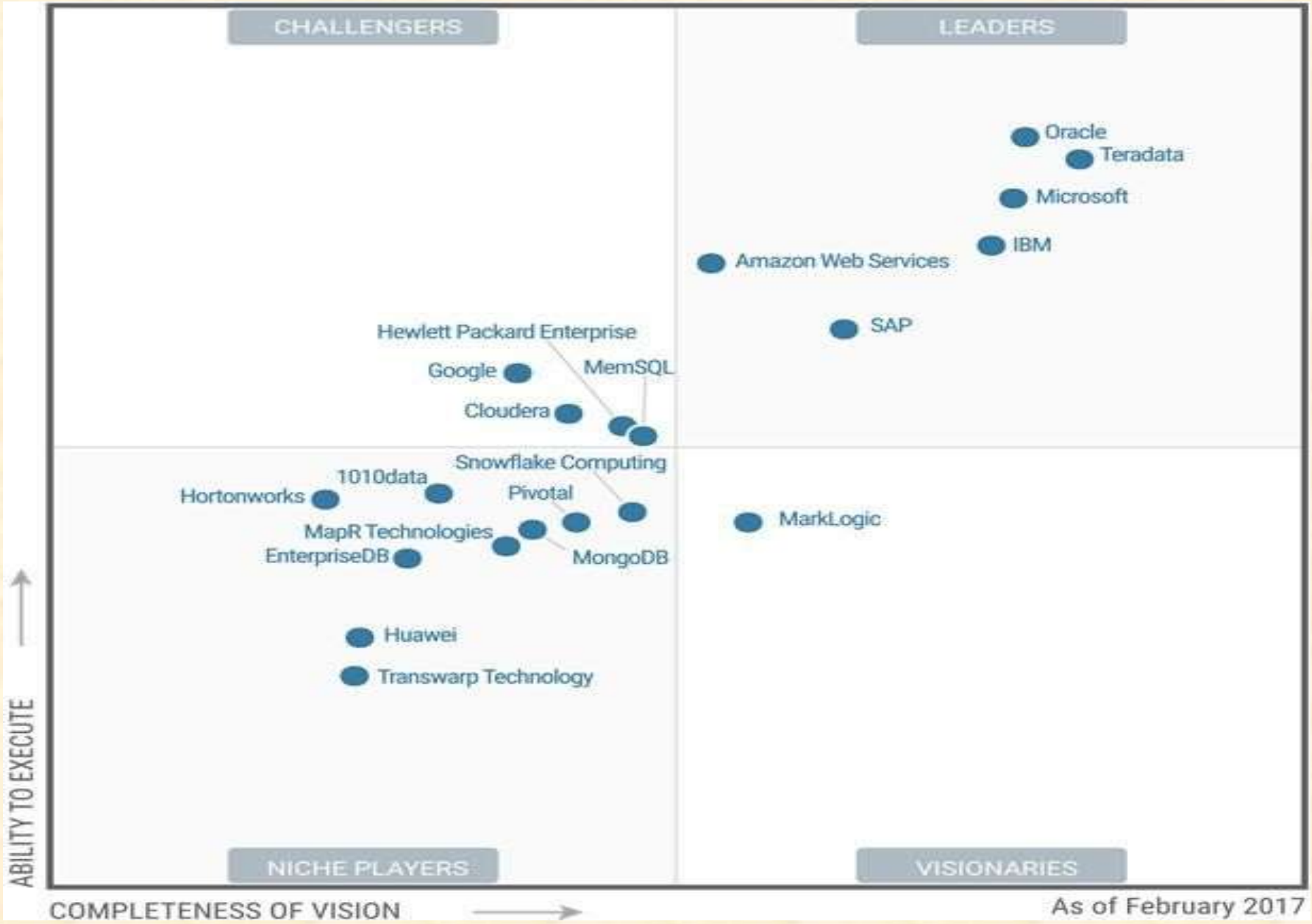
Gartner 2007： 数据仓库提供商



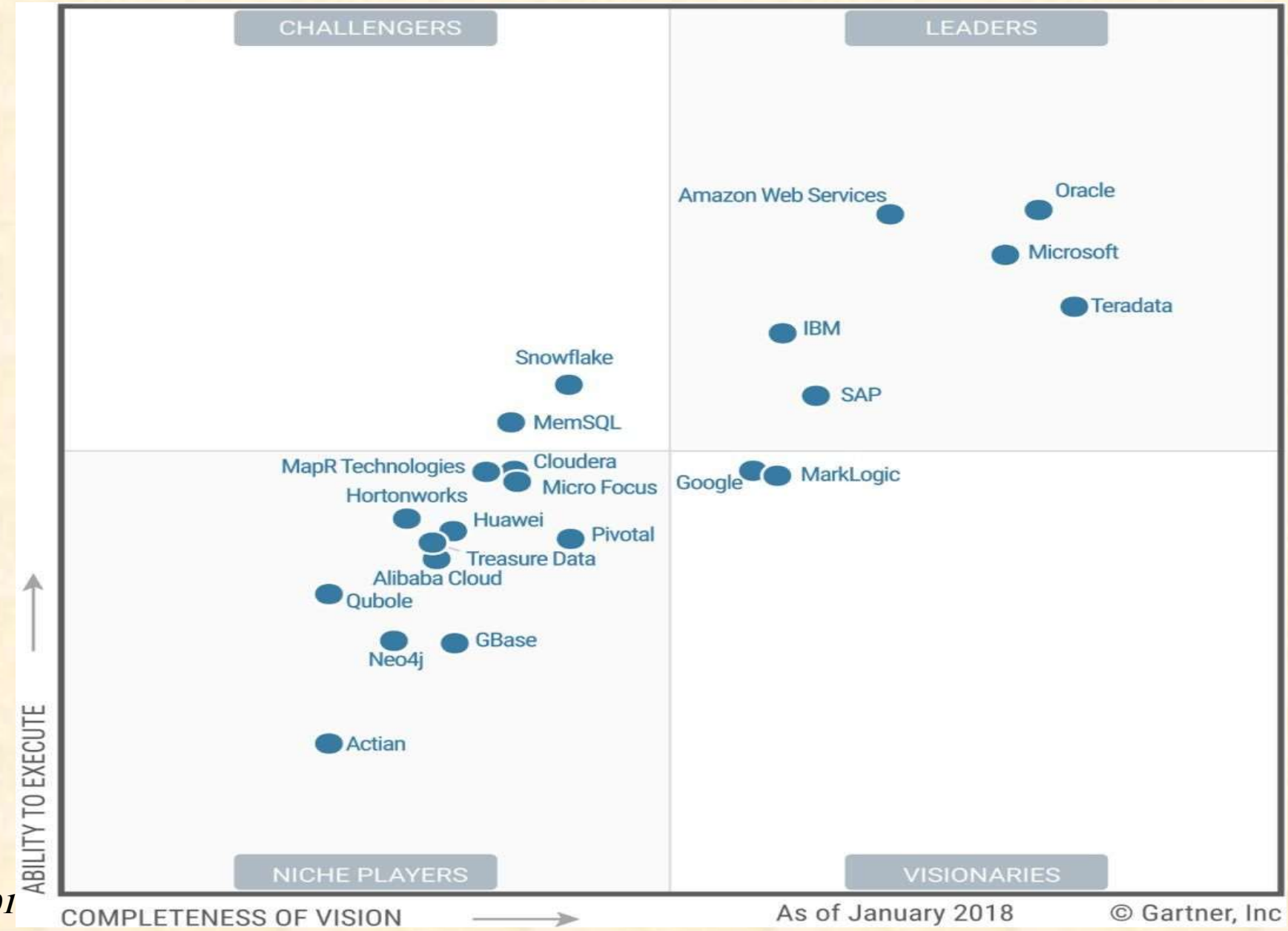
Gartner 2008： 数据仓库提供商



Gartner 2017：面向分析的数据管理解决方案



Gartner 2018：面向分析的数据管理解决方案

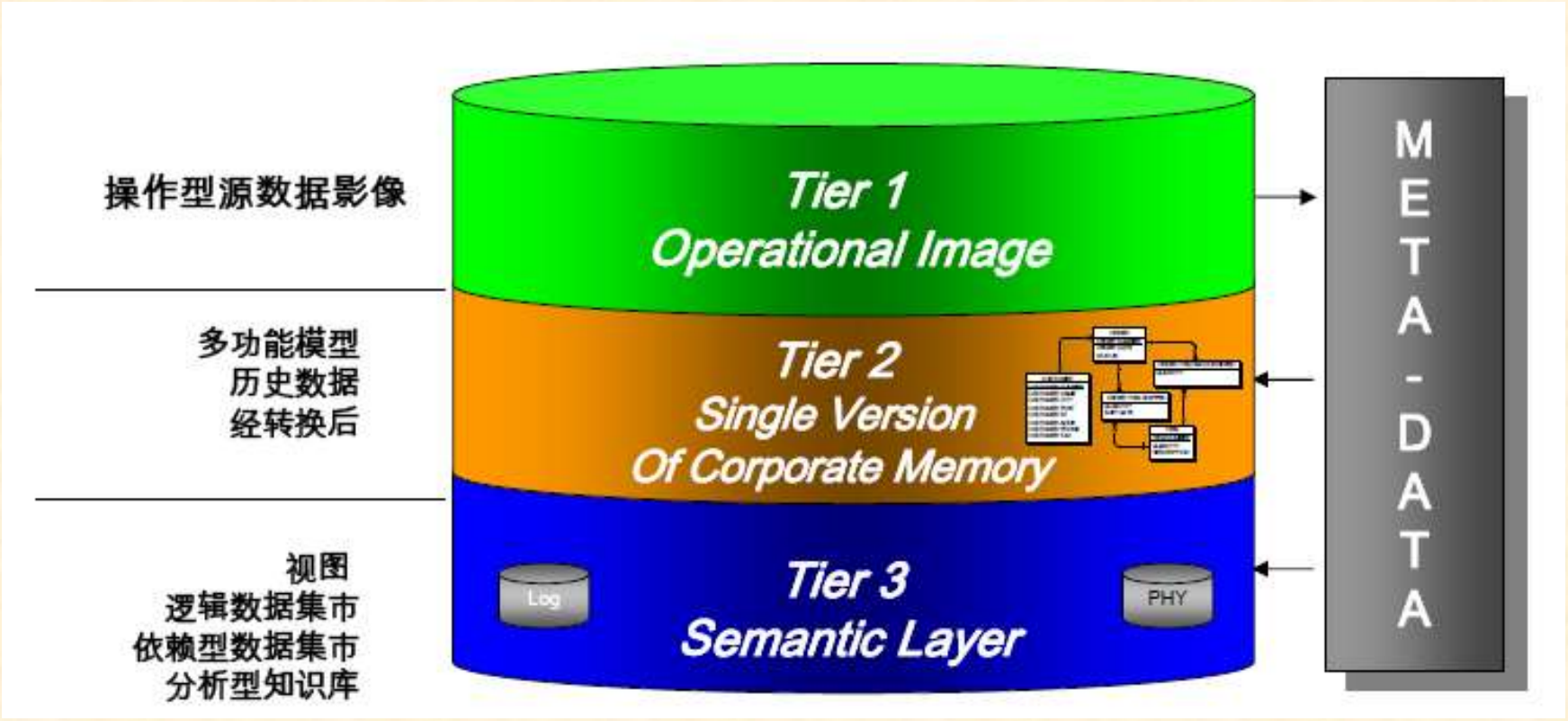


1. 引 言

- 根据功能可分为三大类
  - 解决特定功能的产品
    - 主要包括Teradata的数据仓库解决方案
  - 提供部分解决方案的产品
    - Oracle、IBM、Sybase、Informix、NCR、Microsoft及SAS等公司的数据仓库解决方案
  - 提供全面解决方案的产品
    - CA

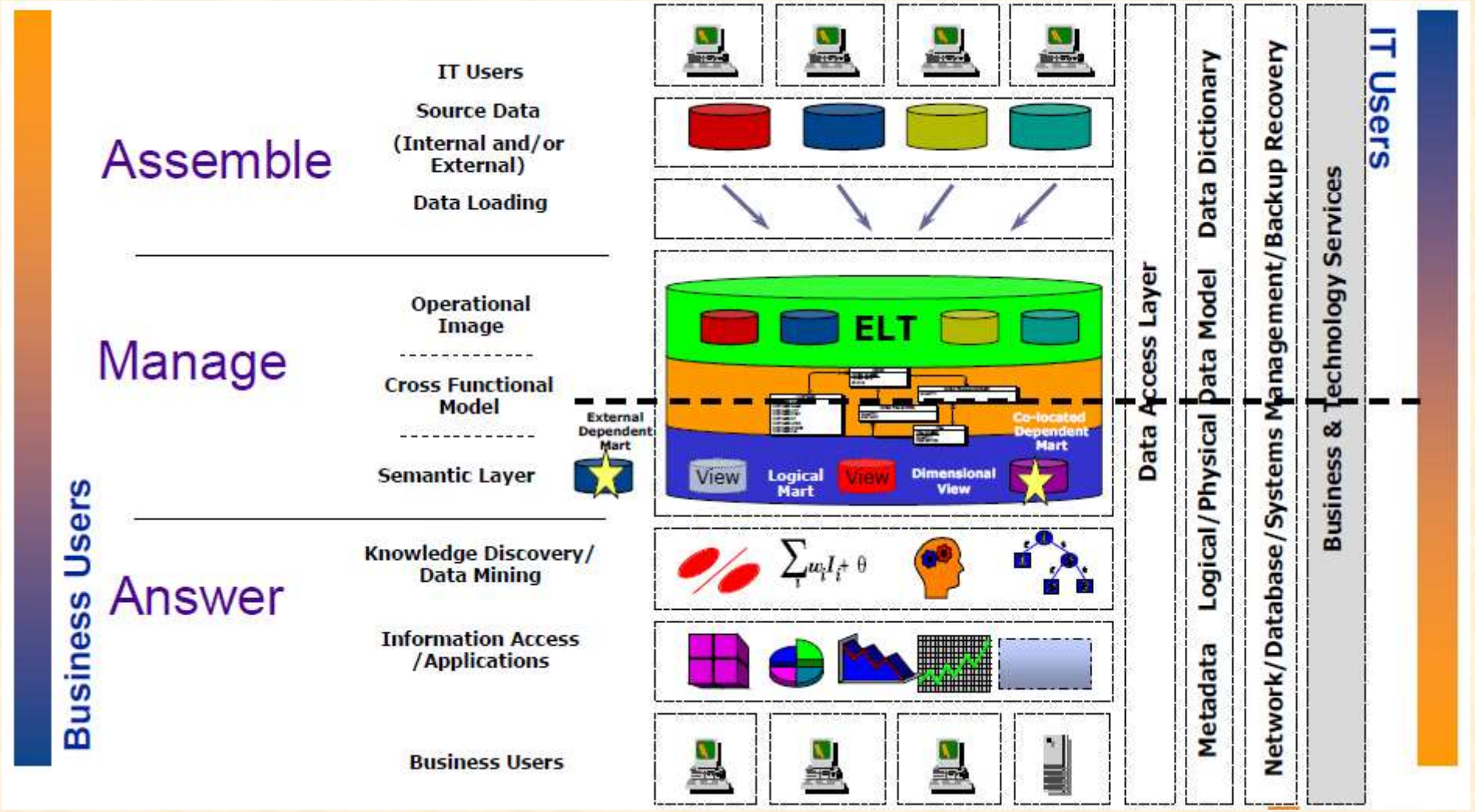
1. 引 言

❑ Teradata EDW 逻辑架构



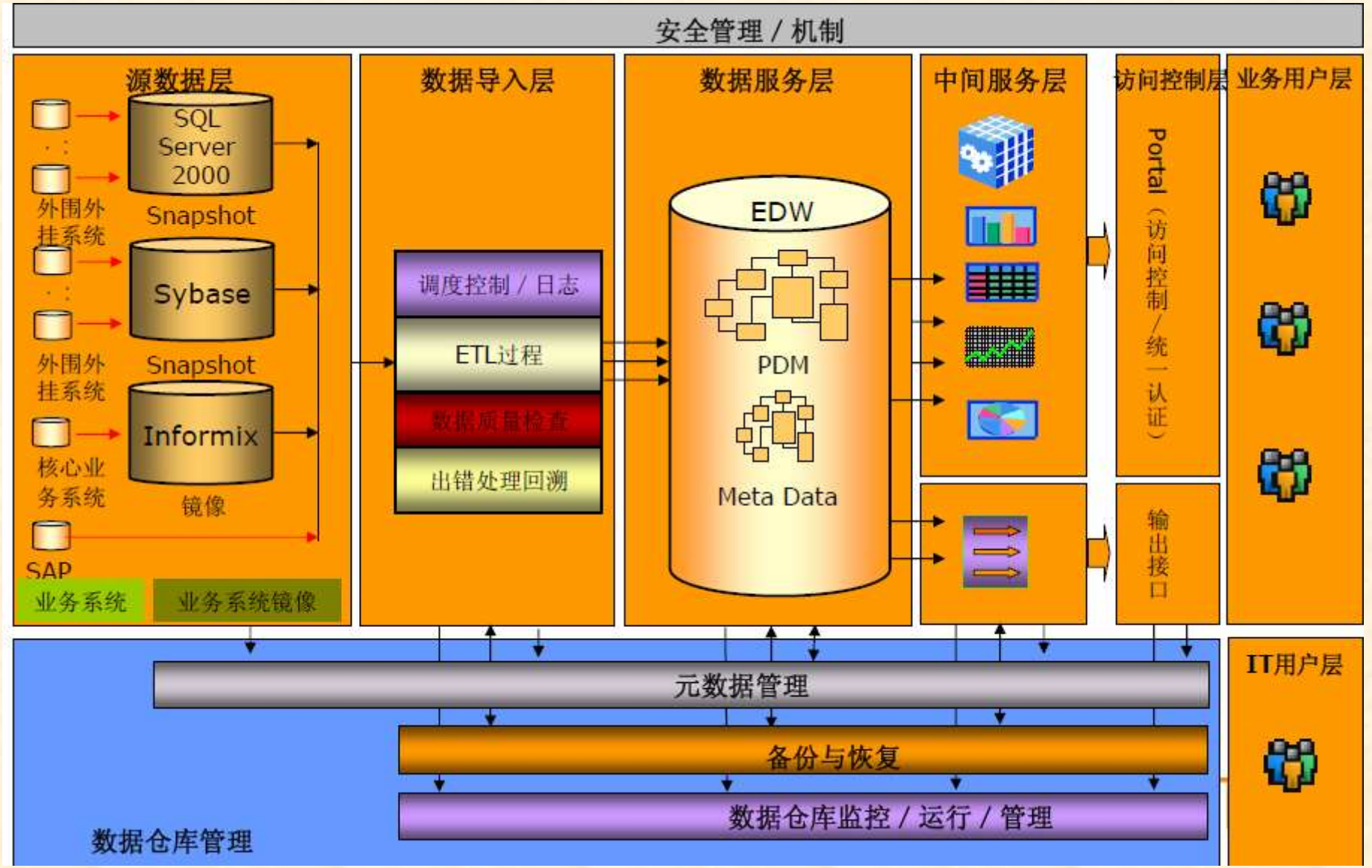
1. 引 言

❑ Teradata EDW 应用架构

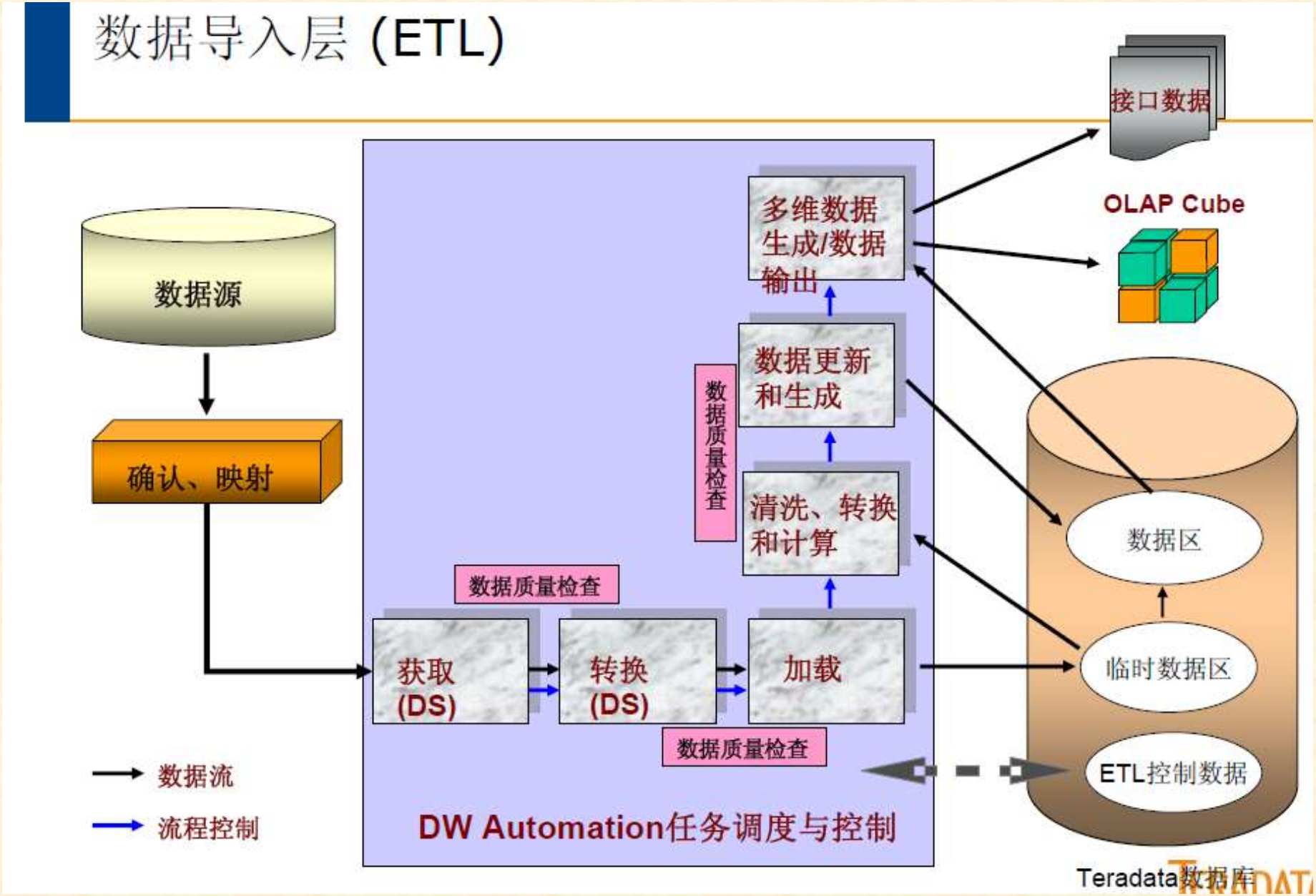


1. 引 言

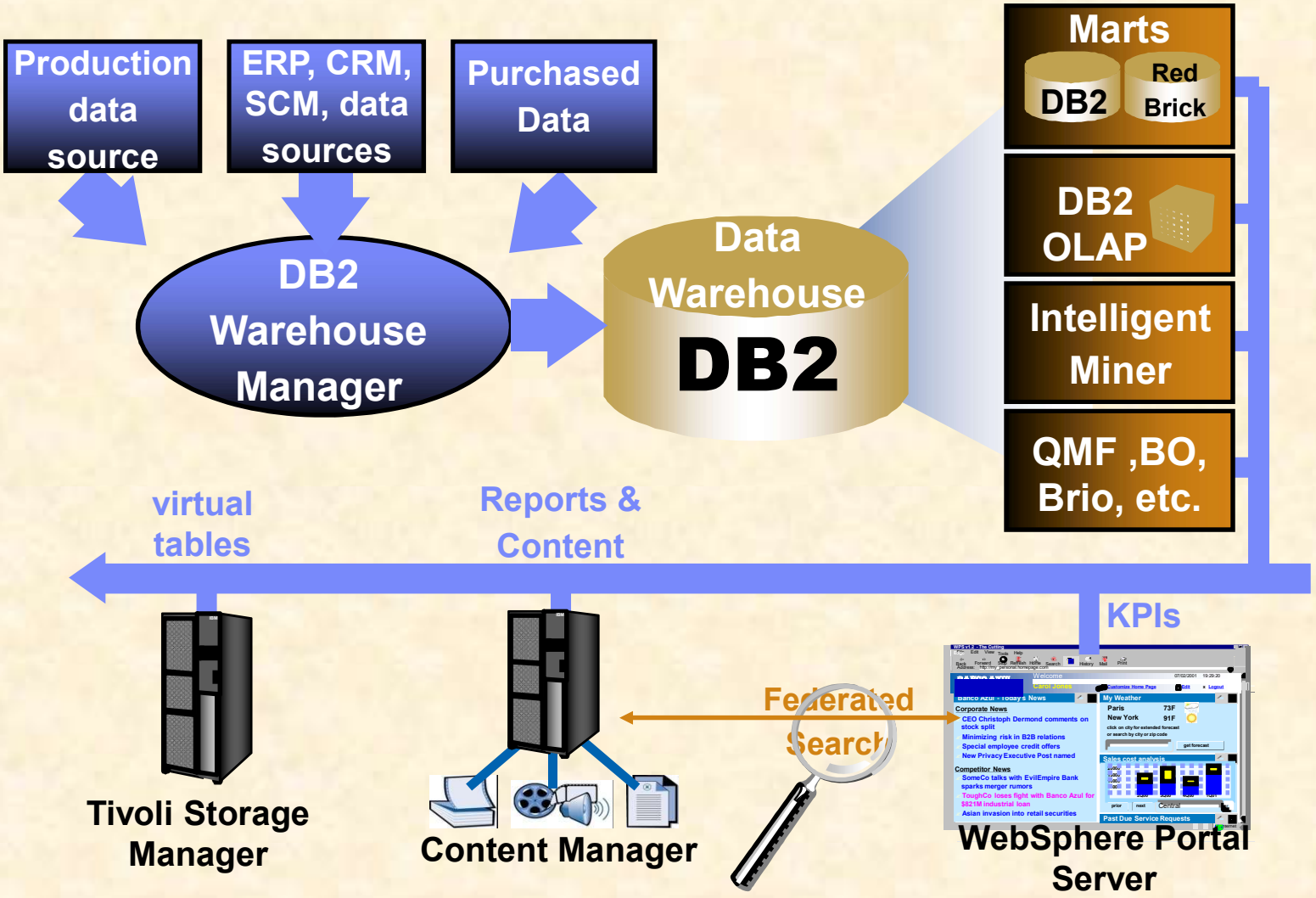
❑ Teradata EDW技术体系架构



1. 引 言


数据导入层 (ETL)
接口数据
OLAP Cube
数据源
确认、映射
多维数据生成/数据输出
数据更新和生成
数据质量检查
清洗、转换和计算
数据质量检查
获取 (DS)
转换 (DS)
加载
数据质量检查
数据区
临时数据区
ETL控制数据
数据流
流程控制
DW Automation任务调度与控制
Teradata数据库

❑ IBM DB2数据仓库技术体系架构



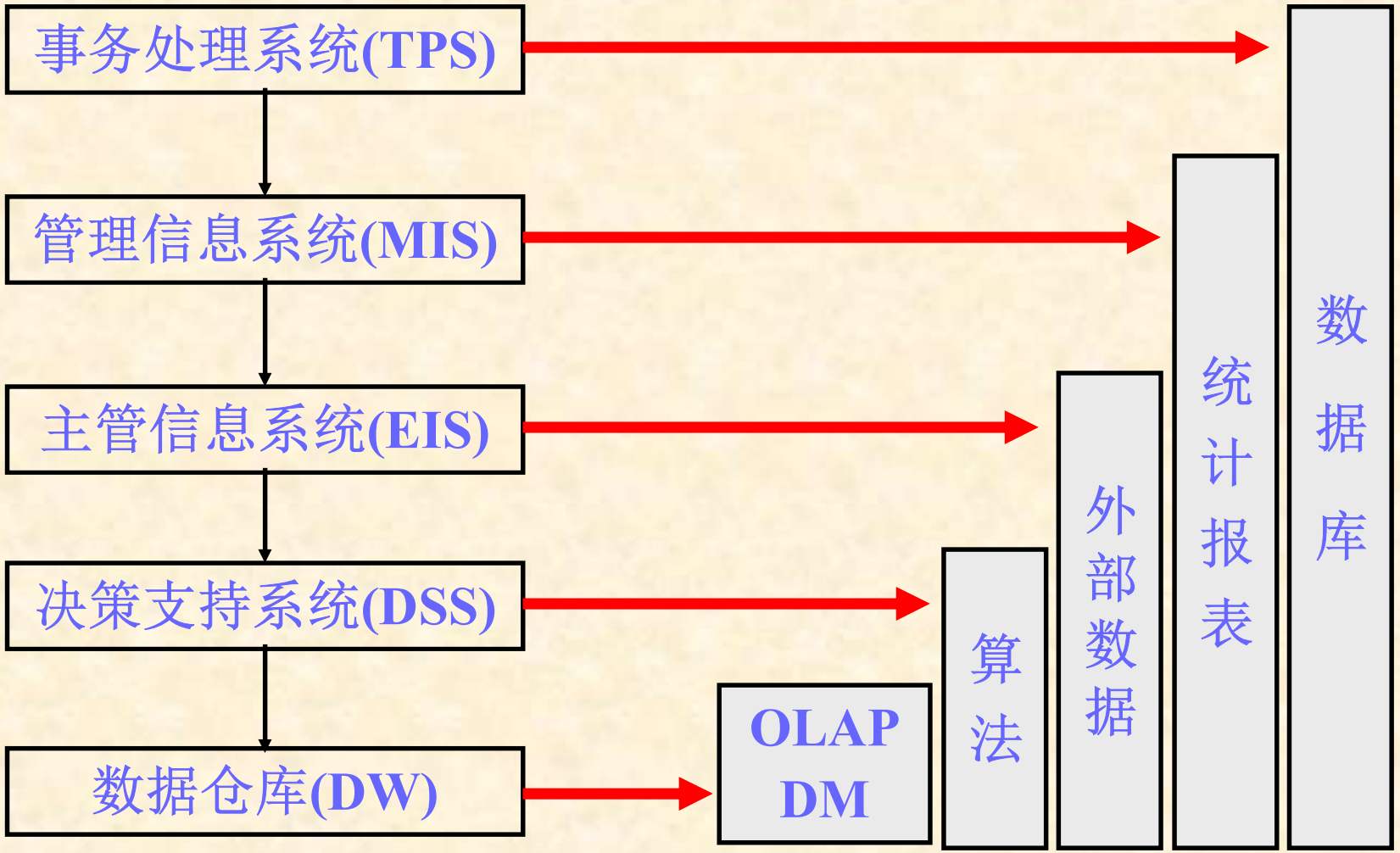
数据仓库

1. 引　言
2. 从数据库到数据仓库　√
3. 数据分析与数据仓库
4. 数据仓库的四大特色
5. 数据仓库的基本结构
6. 数据仓库的设计
7. 联机分析处理 **(OLAP)**
8. 多维建模 **(Dimensional Modeling)**
9. 数据仓库的应用

2. 从数据库到数据仓库

- 数据库技术在事务处理应用中取得了巨大的成功
  - 传统的数据库技术作为数据管理手段，主要用于联机事务处理(OLTP,On-Line Transaction Process)，数据库中保存的是大量的日常业务数据。
  - 在数据共享、数据与应用程序的独立性、维护数据的一致性与完整性、数据的安全保密性等方面提供了有效的手段。
  - 用户对数据信息的需求也越来越复杂，涉及的功能已不仅仅是简单地对少量记录的联机事务处理（OLTP），而是要对多张表中的千万条历史记录进行数据分析和信息综合,即数据分析处理（OLAP）。
  - 因此如何从大量的历史数据中提取有用的知识，已经成为每个企业都要面对的迫切问题。

2. 从数据库到数据仓库



信息系统的发展历史

2. 从数据库到数据仓库

- 操作型处理和分析型处理

  - 随着市场竞争的加剧、企业需求的发展以及数据量的不断增大，数据处理被划分为两大类：

    - 操作型处理

    - 分析型处理

2. 从数据库到数据仓库

- 操作型处理

  - 也叫事务处理，是指对数据库的日常联机访问操作，通常是对一个或一组记录的查询和修改，主要是为企业特定的应用服务，所以也叫联机事务处理。
    - **On-Line Transaction Processing, 简称 OLTP**

  - **OLTP的特点**
    - **a)** 通常仅仅是对一个或一组记录的查询或修改；
    - **b)** 执行频率高；
    - **c)** 人们关心的是处理的响应时间、数据的安全性和完整性等指标。

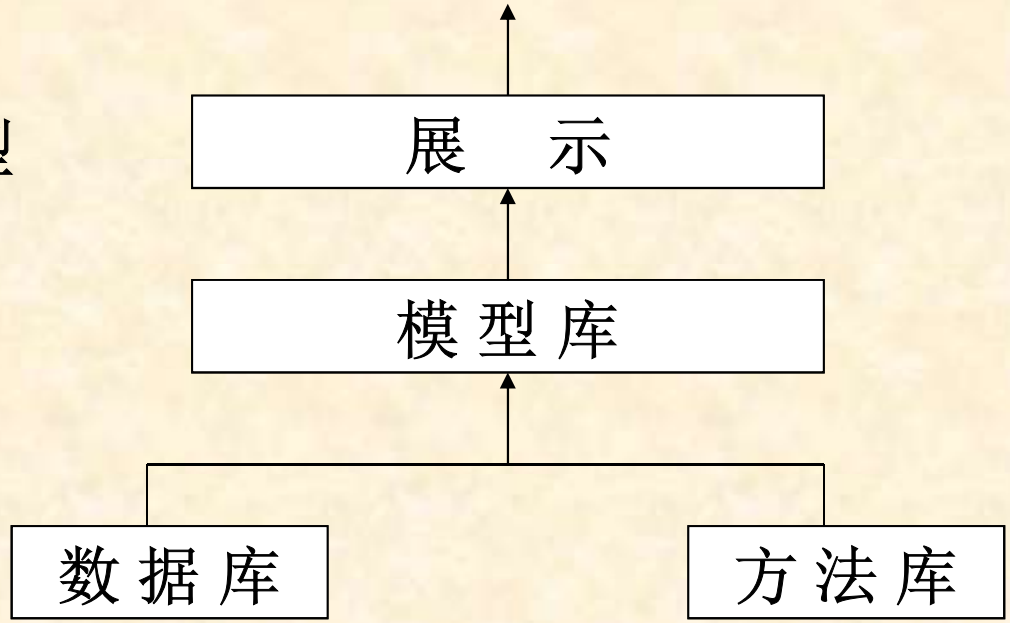
2. 从数据库到数据仓库

- 分析型处理
  - 也叫做信息型处理，主要用于企业管理人员的决策分析，为制订企业的未来经营管理计划提供辅助决策信息。
  - 数据访问特点
    - a) 需要对大量的事务型数据进行统计、归纳和分析；
    - b) 经常访问大量的历史数据；
    - c) 执行频率和对响应时间的要求都不高。
  - 典型的分析型处理系统
    - ➢ **决策支持系统 (DSS --Decision Support System)**

实例： 体育用品公司的数据分析

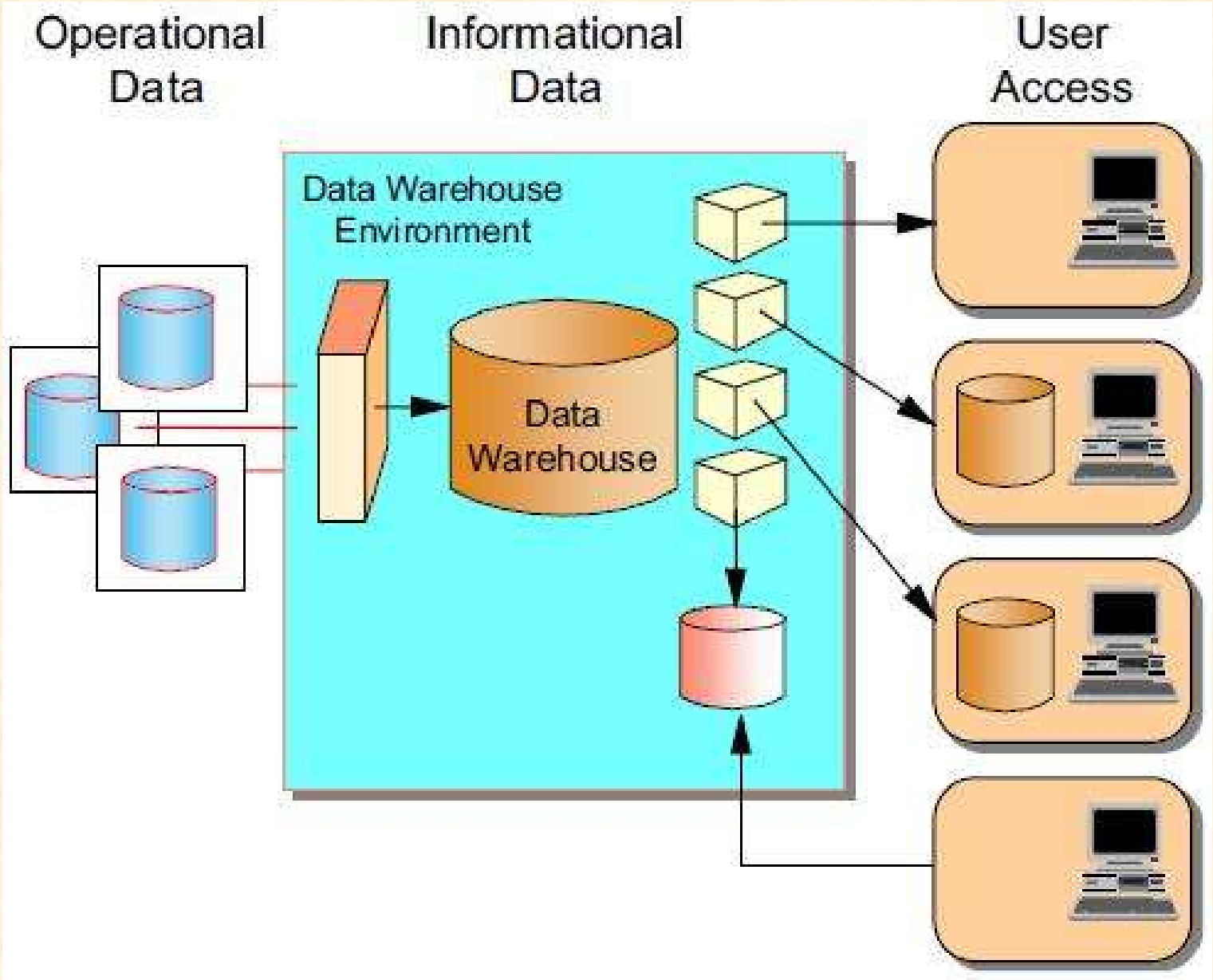
2. 从数据库到数据仓库

- 分析型处理 - 决策支持系统
  - 决策支持系统是上世纪70年代兴起的一种计算机应用技术，用于帮助企业领导作辅助性决策。
  - 传统的决策支持系统由三个组成部分（其结构模型如右下图）
    - 数据
    - 算法与模型
    - 展示



2. 从数据库到数据仓库

- 操作型处理和分析型处理
  - 适于决策的数据通常包括两个部分：
    - 企业的内部数据
      - 如企业财务状况、产品销售情况、客户特征等
    - 企业的外部数据
      - 如企业产品的市场占有率、供应商状况、客户的偏好、市场外部环境的变化等
  - 如何快速地从数据中提取有用信息，制定市场策略，以便对市场机会做出及时灵活的反应，成为企业在信息时代的市场竞争中立于不败之地的关键所在。


Operational Data
Informational Data
User Access
Data Warehouse Environment
Data Warehouse

2. 从数据库到数据仓库

- 在事务处理环境下进行分析处理的不足
  - 为了进行分析型数据处理，人们在关系数据库中放宽了对冗余的限制，引入了统计及综合数据，在事务处理环境下建立了传统的DSS；
    - 但这些统计、综合数据的应用逻辑却是分散杂乱的、非系统化的，因此分析功能有限，不灵活，响应慢，维护困难。
  - 数据库技术一直力图使自己能够胜任从事务处理、批处理到分析处理的各种类型的处理任务；但是作为数据管理手段的数据库技术，尽管在事务处理方面取得了巨大的成功，但它对分析处理的支持却一直不能令人满意。
    - ***数据只为员工服务，不为老板服务！***

2. 从数据库到数据仓库

❑ 事务处理环境不适宜DSS应用的原因

– 在传统的以数据库为核心的事务处理环境中不适宜建立**DSS**等分析型应用，其原因主要有以下六条：

- ➢事务处理和分析处理的性能特性不同
- ➢数据集成问题
- ➢数据的动态集成问题
- ➢历史数据问题
- ➢数据的综合问题
- ➢数据的访问问题

2. 从数据库到数据仓库

（1） 事务处理和分析处理的性能特性不同

- 事务处理
  - 用户每次操作处理的时间短，存取数据量小，但操作频率高，并发程度大。

- 分析处理
  - 每次分析可能需要连续运行很长的时间，存取数据量大，但很少做这样的分析处理，也没有并发执行的要求。

2. 从数据库到数据仓库

（2）数据集成问题

– 事务处理一般只需要与本部门业务有关的当前细节数据，而对整个企业范围内的集成应用考虑很少，这就造成大部分企业内部的数据是分散而非集成的。

– 造成上述状况的原因是多方面的：

➢ 事务处理应用的分
➢ 数据不一致问题
➢ 缺少分析所需要的

- 数据类型、单位的不一致性
- 同名异义、同义异名现象
- 因数据的重复抽取而带来的数据不一致性

2. 从数据库到数据仓库

（2）数据集成问题 (cont.)

- 分析处理
  - 全面而正确的数据是有效的分析和决策的首要前提。
  - **DSS**需要集成的数据，包括整个企业内部各部门的相关数据，以及企业外部、竞争对手等处的相关数据。因此用于分析处理的数据可能来自多种不同的数据源，包括：
    - 同构/异构数据库
    - 文件系统
    - **Internet**
    - 外部的用户数据

2. 从数据库到数据仓库

（2）数据集成问题 (cont.)

- 对于需要集成数据的DSS应用来说，在应用程序中对事务处理环境中的这些纷繁复杂的数据进行集成将带来下述问题：
  - 大大加重程序员的负担
  - 重复计算
  - 极低的分析处理效率

2. 从数据库到数据仓库

（3）数据的动态集成问题

- 静态集成
  - ➢对所需数据进行一次集成，以后就不再发生变化
- 动态集成
  - ➢对集成后的数据进行周期性刷新

■ 在采用静态集成策略时，如果数据源中的数据发生了变化，那么这些变化就不能反映给决策者，导致决策使用的是过时的数据。因此集成数据必须以一定的周期进行刷新（*即采用动态集成策略*），但传统的事务处理环境并不具备动态集成的能力。

2. 从数据库到数据仓库

（4）历史数据问题

- 事务处理
  - 一般只需要当前数据，在数据库中一般也只存储短期数据 (3-6个月)，且不同数据的保存期限也不一样。
    - 数据库中的过时数据（即历史数据）虽然也能通过数据转储等方式保存下来，但往往被束之高阁，未能得到充分利用。
- 分析处理
  - 更看重历史数据 (5-10年)，可以通过对大量历史数据的详细分析来把握企业的发展趋势。

- 历史数据对于事务处理作用不大，但对于决策分析而言，如果没有历史数据的支撑，就变成了“无源之水”、“无本之木”。

2. 从数据库到数据仓库

（5）数据的综合问题

- 事务处理需要的是当前的细节性操作数据，而分析处理需要的往往是大量的总结性分析型数据，而非数据库中的细节性操作型数据。

- 事务处理系统中积累的是大量的细节数据，而DSS并不对这些细节数据进行分析，其原因是：
  - 细节数据量太大，影响处理效率
  - 不利于分析人员将注意力集中于有用的信息上

2. 从数据库到数据仓库

（5）数据的综合问题（cont.）

- 这就是常说的数据库中“数据丰富、信息贫困”现象。
- 因此，在分析前往往需要对细节数据进行不同程度的综合，传统的事务处理系统不具备这种综合能力，而且在数据库系统中，这种综合还往往因为是一种数据冗余而被限制。

2. 从数据库到数据仓库

（6）数据的访问问题

- 事务处理
  - ➢需要提供多种不同类型的数据访问操作
  - ➢对于需要修改的数据必须实时‘更新’数据库
- 分析处理
  - ➢数据的访问操作以‘读’操作为主
  - ➢不需要实时的‘更新’操作，但需要定时‘刷新’

2. 从数据库到数据仓库

- *事务处理环境不适宜DSS应用的原因（cont.）*
  - 综上所述，在事务处理环境中直接构建分析处理应用是不合适的，要提高分析处理和决策支持的效率和有效性，必须将**分析型处理及其所需的综合性数据从传统的事务型处理和细节性数据中分离出来**，按照DSS的需要重新进行组织，建立单独的分析处理环境，**数据仓库**正是为建立这种新的分析处理环境而出现的一种数据管理技术。

2. 从数据库到数据仓库

- 将数据仓库与数据库分离开来的好处：
  1）提高两个系统的性能
  2）提高操作型数据库的事务吞吐量
  3）两个系统中数据的结构、内容和用法的不同

- 建立数据仓库的目的并不是要代替传统的事务处理系统（数据库），而是为了适应因商业经营行为的改变和市场竞争程度的加剧而进行的DSS的需要。

- 目前，数据仓库技术已成为企业信息集成和辅助决策应用的关键技术之一。

2. 从数据库到数据仓库

- *数据仓库发展简史*
  - **1981：NCR公司为Walmart 建立了第一个数据仓库，总容量超过101TB（OLTP：10年的会计文档<1TB)**
  - **1983：Teradata公司利用并行处理技术为美国富国银行（Wells Fargo Bank）建立了第一个DSS**
  - **1988：IBM研究员Barry Devlin和Paul Murphy提出术语：数据仓库(Data Warehouse)**
  - **1992：Bill Inmon出版《Building the Data Warehouse》，第一次给出了数据仓库的清晰定义和操作性极强的指导意见，开始得以大规模应用（数据仓库之父）**
  - **1993：斯坦福博士Kimball提出由下而上，从部门到企业的数据仓库构建方式，从易到难，得到了长足的发展**
  - 数据仓库已成为“数据 ➡知识➡利润”的基础核心技术。

数据仓库

1. 引　言
2. 从数据库到数据仓库
3. 数据分析与数据仓库　√
4. 数据仓库的四大特色
5. 数据仓库的基本结构
6. 数据仓库的设计
7. 联机分析处理(OLAP)
8. 多维建模 (Dimensional Modeling)
9. 数据仓库的应用

3. 数据分析与数据仓库

– 在现代计算机信息系统中，数据的作用有两个方面：事务处理和分析处理，不同的用户（处理）需要不同的数据信息。

➢操作型数据

- 事务处理所需要的细节性的数据，是面向企业员工的日常业务处理过程的，通常由数据库管理系统来负责其存储与管理。

➢分析型数据

- 分析处理所需的综合性数据，是面向企业管理人员的决策需要的。

3. 数据分析与数据仓库

特 性	操作型数据（DB）	分析型数据（DW）
定位	面向应用的事务处理	面向主题的数据分析
DB设计	E-R模型	星型/雪花模型，数据立方体
数据	当前的、最新的	历史的，具有时间跨度
汇总	原始的，细节的	集成的，一致的
视图	详细的，关系的	总体的，多维的
操作类型	读/写（易变的）	读（稳定的）
存取请求	可预知的	事先未知的
访问记录	一次操作少量记录	一次操作大量记录
数据规模	MB ~ GB	TB ~ PB
工作单位	短的，简单事务	复杂查询
性能要求	对性能要求高	对性能要求较宽松

数据仓库与数据库系统的比较

3. 数据分析与数据仓库

- 数据仓库的定义
  - W. H. Inmon 在《建立数据仓库》一书中，对数据仓库的定义为：
    - 数据仓库就是一个面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集合，用于支持经营管理过程中的决策制定。
  - Tim. Shelter（Informix公司负责研究与开发的副总裁）
    - 数据仓库将分布在企业网络中不同信息岛上的商业数据集成到一起，存贮在一个单一的集成关系型数据库中。利用这种集成信息，可方便用户对信息的访问，更可使决策人员对一段时间内的历史数据进行分析，研究事物发展走势。

3. 数据分析与数据仓库

- 根据上面的定义，我们可以从两个层面上来理解数据仓库的概念：

1) 首先，数据仓库用于支持用户的决策行为，面向分析型数据处理，它不同于企业现有的操作型数据库；

2) 其次，数据仓库是对多个异构的数据源的有效集成，然后按照主题进行了重组，并包含历史数据，而且存放在数据仓库中的数据一般不再修改。

3. 数据分析与数据仓库

- 数据仓库的定义
- Data warehouse is a subject oriented, integrated,non-volatile and time variant collection of data in support of management's decision — *[Inmon,1996]*.
- Data warehouse is a set of methods, techniques,and tools that may be leveraged together to produce a vehicle that delivers data to end-users on an integrated platform — *[Ladley,1997]*.
- Data warehouse is a process of crating, maintaining,and using a decision-support infrastructure — *[Appleton,1995] [Haley,1997] [Gardner 1998].*

3. 数据分析与数据仓库

- 数据仓库的特征
  - 面向主题
  - 集成
  - 稳定性
  - 随时间而变化（时间维）

3. 数据分析与数据仓库

- 数据仓库中的基本概念
  - Data Mart（数据集市）
    - 小型的，面向部门或工作组级数据仓库。
  - Operation Data Store（操作数据存储）
    - 能支持企业日常的全局应用的数据集合，是不同于DB的一种新的数据环境，是DW扩展后得到的一个混合形式。
    - 四个基本特点：
      - 面向主题的、集成的、可变的、当前或接近当前的

3. 数据分析与数据仓库

- 数据仓库： 用于支持管理决策(Decision Making)。

- 实际中存在需要全部数据但又变化较快的情形
  - 例如：订单类数据，可能需要频繁进行全量更新（因为：同一个订单状态随着时间会变化，比如今天买了，明天退货了）。

- ODS：用于支持企业对于即时性的、操作性的、集成的全体信息的需求。
  - ODS：短期的实时的数据，供产品或者运营人员日常使用，而数据仓库是供战略决策使用的数据；
  - ODS是可以更新的数据，数据仓库是基本不更新的反应历史变化的数据；
  - ……

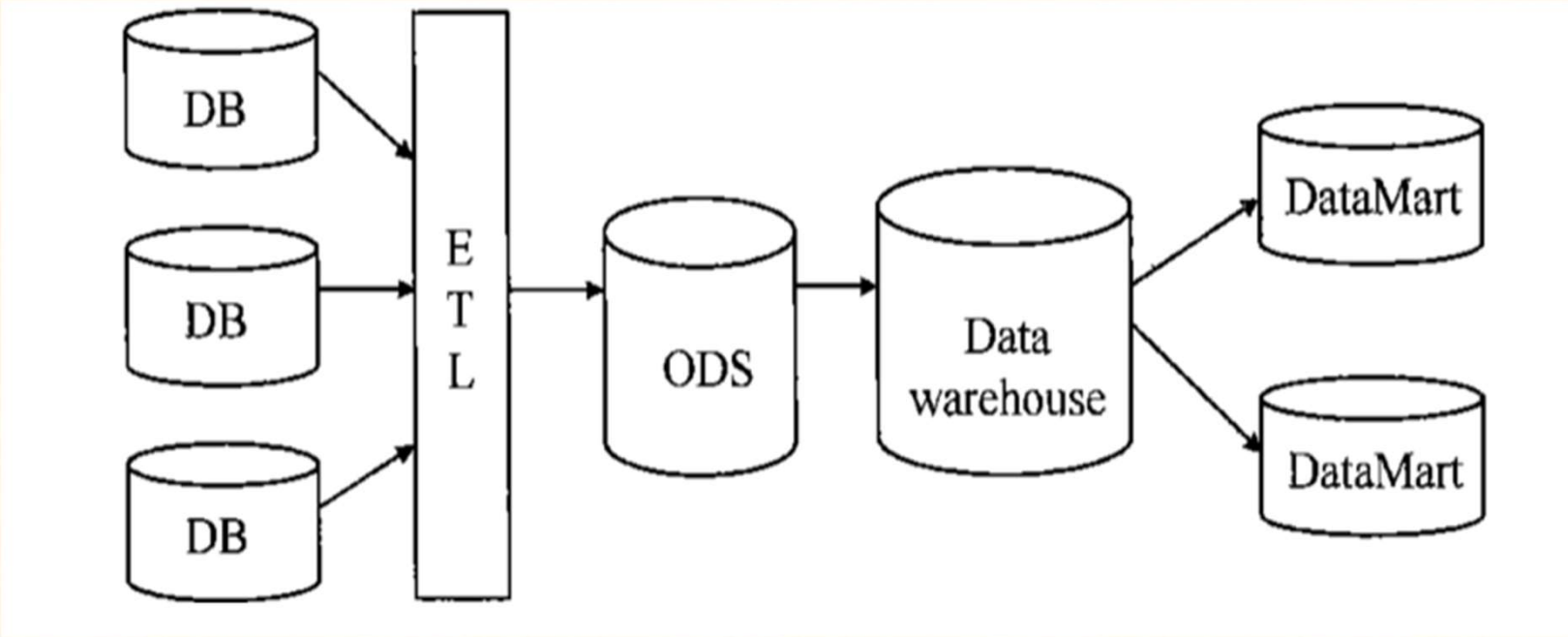
概要比较：

分类	ODS系统	DW系统
目的	接近实时监控	决策支持
共同点	整合数据	整合数据
	面向主题	面向主题
不同点	动态数据（延迟>1秒）	静态数据（延迟>24小时）
	当前数据	历史数据
	细节化数据	概括性数据

实施方案比较：

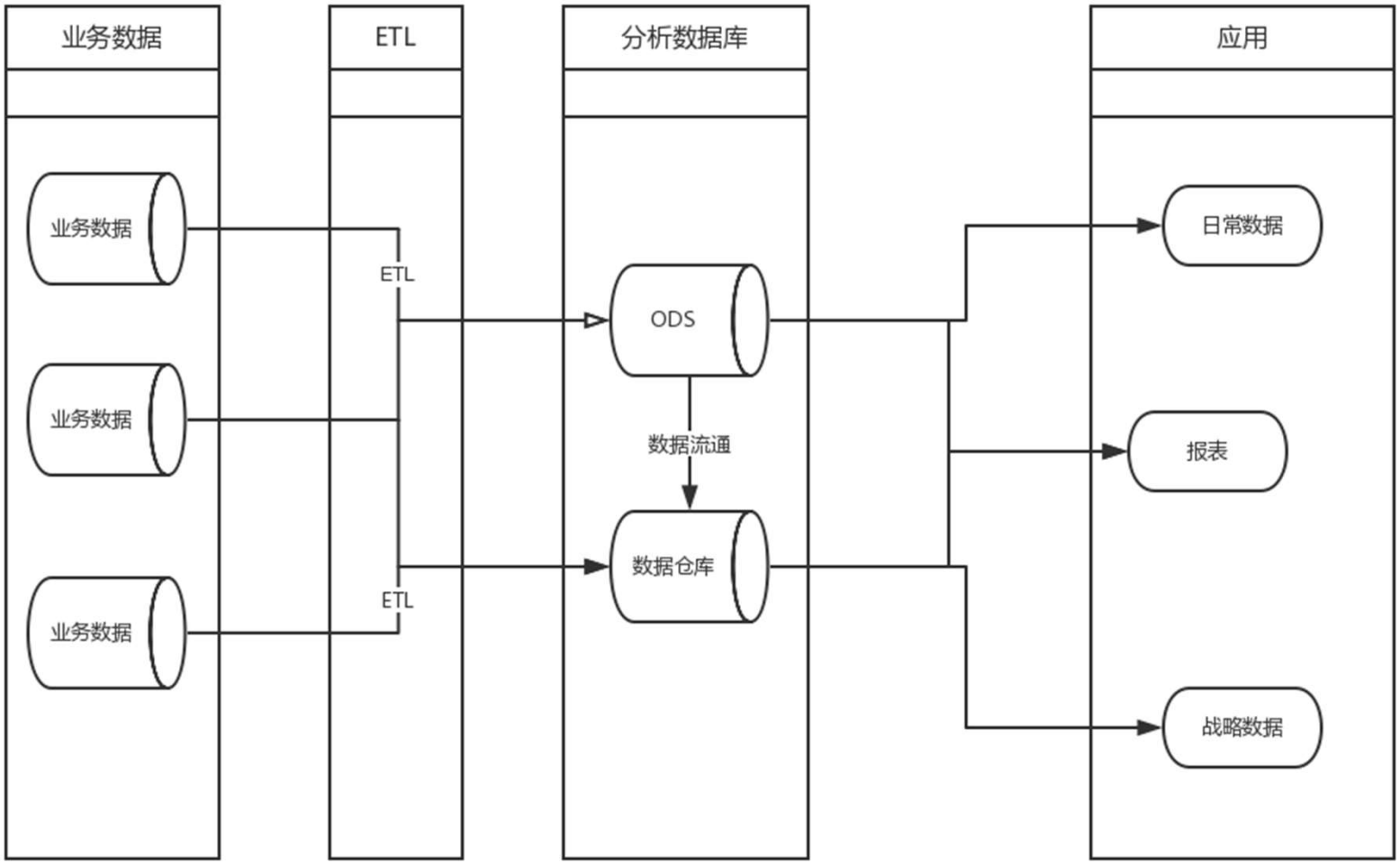
实施方案	实施结果	优势	劣势
DW	企业能够分析DW中的历史数据，进行中远期的规划	可以解决企业的决策需求	不能满足企业的实时监控和实时业务需求
ODS	企业能够把握ODS中的当前综合数据对企业的及时运行情况随时掌控	,可以满足企业的实时监控和实时业务需求	不能满足企业的中远期决策需求

DW+ODS架构1



- 优点：**ODS**的数据与数据仓库的数据高度统一；开发成本低，开发一次并应用到**ODS**即可。
- 缺点：数据仓库需要的所有数据都需要经由**ODS**，**ODS**的灵活性必然受到影响，不利于扩展，系统的灵活性差

DW+ODS架构2



- 便于扩展，**ODS**和数据仓库实现优势互补

3. 数据分析与数据仓库

- **ETL**
  - **ETL（Extract/Transformation/Load）**
  - 数据抽取、转换、装载工具
- 元数据
  - 关于数据的数据，用于构造、维持、管理、和使用数据仓库，在数据仓库中尤为重要。
- 粒度
  - 数据仓库的数据单位中保存数据的细化或综合程度的级别。
  - 细化程度越高，粒度越小。
- 分割
  - 数据分散到各自的物理单元中去，它们能独立地处理。

3. 数据分析与数据仓库

❑ 数据仓库与DSS

- 三类分析工具可用于决策支持
- **OLAP**：能够支持涉及分组和聚集查询，并能够对各种复杂的布尔条件、统计函数和时间序列分析提供支持的系统。
- 支持传统**SQL**查询的**DBMS**，但为有效地执行**OLAP**查询而进行了特殊的设计。这些系统可以看作是为决策支持应用进行了优化的关系数据库系统。
- 数据挖掘

数据仓库

1. 引　言
2. 从数据库到数据仓库
3. 数据分析与数据仓库
4. 数据仓库的四大特色　√
5. 数据仓库的基本结构
6. 数据仓库的设计
7. 联机分析处理(OLAP)
8. 多维建模 (Dimensional Modeling)
9. 数据仓库的应用

4. 数据仓库的四大特色

- 数据仓库的四个特色：
  - 面向主题
  - 集成
  - 相对稳定（或：不可更新，非易失）
  - 随时间不断变化(时间维，跨度相对较大)

4. 数据仓库的四大特色

（1）面向主题

– 主题（Subject）

➢主题是用户使用数据仓库进行决策时所关心的重点方面，每一个主题基本对应一个宏观的分析领域。

➢例如：

- **CRM：**“客户”主题
  - 优质客户的挖掘
  - 新客户的发现
  - ……
- **ERP：**“产品”主题
  - 销售管理
  - 产品质量控制
  - 库存管理
  - ……

4. 数据仓库的四大特色

（1）面向主题 (cont.)

- 面向主题
  - 面向主题是指数据仓库内的信息是按主题进行组织的，为按主题进行决策的过程提供信息。
    - 为特定数据分析领域提供的数据与传统数据库中的数据是有不同的。传统数据库中的数据是*原始、基础数据*，而特定分析领域数据则是需要对它们作必要的*抽取、加工与总结*而形成。
  - 数据仓库是面向分析、决策人员的主观要求。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，因此，*数据仓库中的主题有时会因用户主观要求的变化而变化*。

4. 数据仓库的四大特色

（1）面向主题 (cont.)

– 例：一个面向事务处理应用的“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）

4. 数据仓库的四大特色

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）

4. 数据仓库的四大特色

采购子系统：	销售子系统：	库存子系统：	人事子系统：
订单	顾客	领料单	员工
订单细则	销售记录	进料单	部门
供应商		库存	
		库房	

- 上述数据模式基本上是按照企业内部的业务活动及其需要的相关数据来组织数据的存储，没有实现真正的数据与应用分离，其抽象程度也不够高。

4. 数据仓库的四大特色

- 如果按照面向主题的方式进行数据组织，首先应该抽取主题，即按照管理人员的分析要求来确定主题，而与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。

- 在该例中，我们可以抽取出三个不同的主题（即分析对象）及其相关的数据：
  - 商品
  - 供应商
  - 顾客

4. 数据仓库的四大特色

- 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

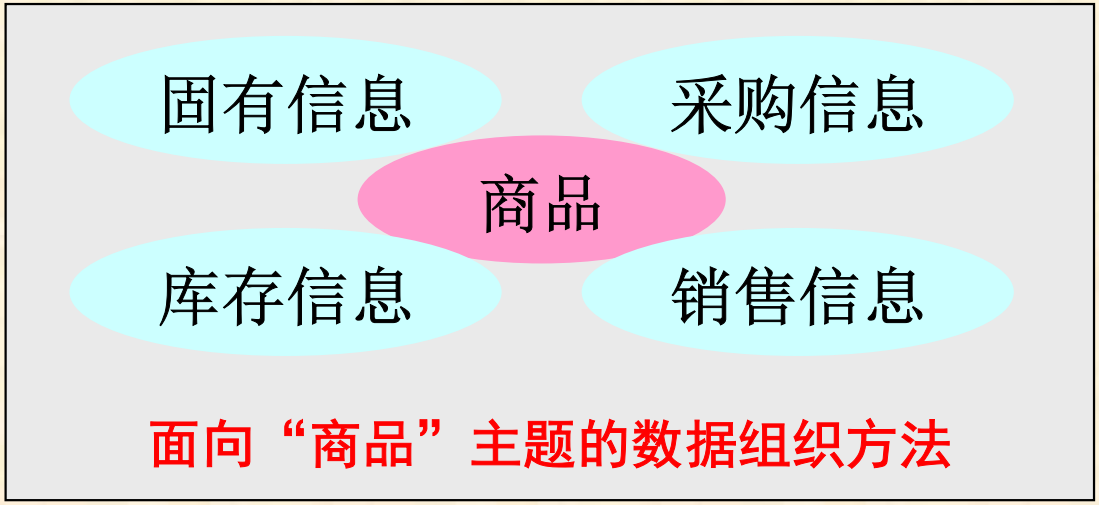
4. 数据仓库的四大特色

- 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

4. 数据仓库的四大特色

（1）面向主题 (cont.)

– 在每个主题中，都包含了有关该主题的所有信息，同时又抛弃了与分析处理无关或不需要的数据，从而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。面向主题的数据组织方式所强调的就是要形成一个这样一致的信息集合。



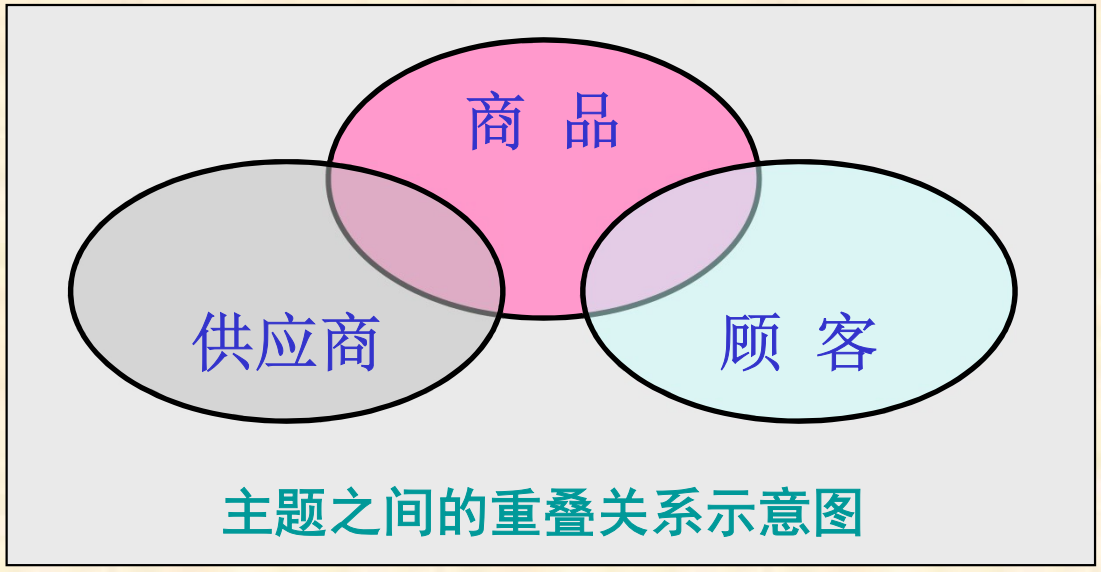
面向“商品”主题的数据组织方法

4. 数据仓库的四大特色

（1）面向主题 (cont.)

– 不同的主题之间也有重叠的内容，但这种重叠的特点是：

- 是逻辑上的，而不是物理存储上的重叠；
- 是部分细节的重叠，而不是统计信息的重叠；
- 可以反映不同主题之间的直接和间接的联系。



主题之间的重叠关系示意图

4. 数据仓库的四大特色

（1）面向主题 (cont.)

– 每个主题所需数据的物理存储：

➢ **多维数据库（MDDB—Multi-Dimensional DataBase）**

- 用多维数组形式存储数据。

➢ **关系数据库**

- 用一组关系来组织数据的存储，同一主题的一组关系都有一个公共的关键字。这是目前实现数据仓库中数据的物理存储的常用方法。
- 在关系中存放的不是细节性的业务数据，而是经过一定程度的综合形成的综合性数据。

4. 数据仓库的四大特色

- 以‘商品’这个主题为例，其公共码键是‘商品号’，其关系存储如下：

- 商品的固有信息
  - 细节数据
    - 商品表（商品号，商品名，类型，颜色，…）
  - 综合数据
    - 商品表1（商品号，商品类别，商品颜色）
    - 商品表2（商品号，价格）
    - ……

4. 数据仓库的四大特色

- 采购信息
  - 细节数据
    - 采购表（商品号，供应商号，供应日期，供应价，…）
  - 综合数据：根据不同的时间段（月、季度、年）来统计商品的采购总量
    - 采购表**H1**（商品号，时间段**1**，采购总量，…）
    - ……
    - 采购表**Hn**（商品号，时间段**n**，采购总量，…）

4. 数据仓库的四大特色

- 销售信息
  - 细节数据
    - 销售表（商品号，顾客号，销售日期，售价，销售量，…）
  - 综合数据：根据不同的时间段（日、周、月、年）统计得到的销售总量
    - 销售表1（商品号，时间段1，销售总量，…）
    - ……
    - 销售表n（商品号，时间段n，销售总量，…）

4. 数据仓库的四大特色

- 库存信息
  - 细节数据
    - 库存表（商品号，库房号，库存量，日期，…）
  - 综合数据：根据不同的时间段（周，月，季度，年）抽样得到的商品库存数量
    - 库存表1（商品号，库房号，库存量，星期，…）
    - ……
    - 库存表n（商品号，库房号，库存量，年份，…）

4. 数据仓库的四大特色

- “面向主题” 总结
  - 一个主题领域的表来源于多个操作型应用（如：客户主题，来源于：定单处理；应收帐目；应付帐目；...）
  - 典型的主题领域：客户；产品；交易；帐目
  - 主题领域以一组相关的表来具体实现
  - 相关的表通过公共的键码联系起来（如：顾客标识号 **Customer ID**）
  - 每个键码都有时间元素（从日期到日期；每月累积；单独日期...）
  - 主题内数据可以存储在不同粒度上（综合级，细节级，多粒度）

4. 数据仓库的四大特色

（2）集成

- 产品的动态成本反映在ERP、CRM、SCM系统里相差很大。
  - 引用ERP和CRM里面的数据：很成功、销量很好的产品
  - SCM数据：采购和物流成本过高，一笔赔钱买卖。
- 来自不同系统的数据基础产生不同的判断：系统并不会去周密地“思考”在自己“职责”之外的事情。
- 后果：给企业的领导提交了相当多顾此失彼的分析报告，导致市场决策上的混乱和失误。
- 把企业的内部数据和外部数据进行有效的集成，形成直观的、易于理解的信息，再进行分析和思考，为企业的各层决策及分析人员使用。
  - 内部数据：业务系统（SCM、ERP、CRM）数据，不同硬件、数据库、网络环境，为不同的业务部门服务。
  - 外部数据是市场信息和外部竞争对手的信息。

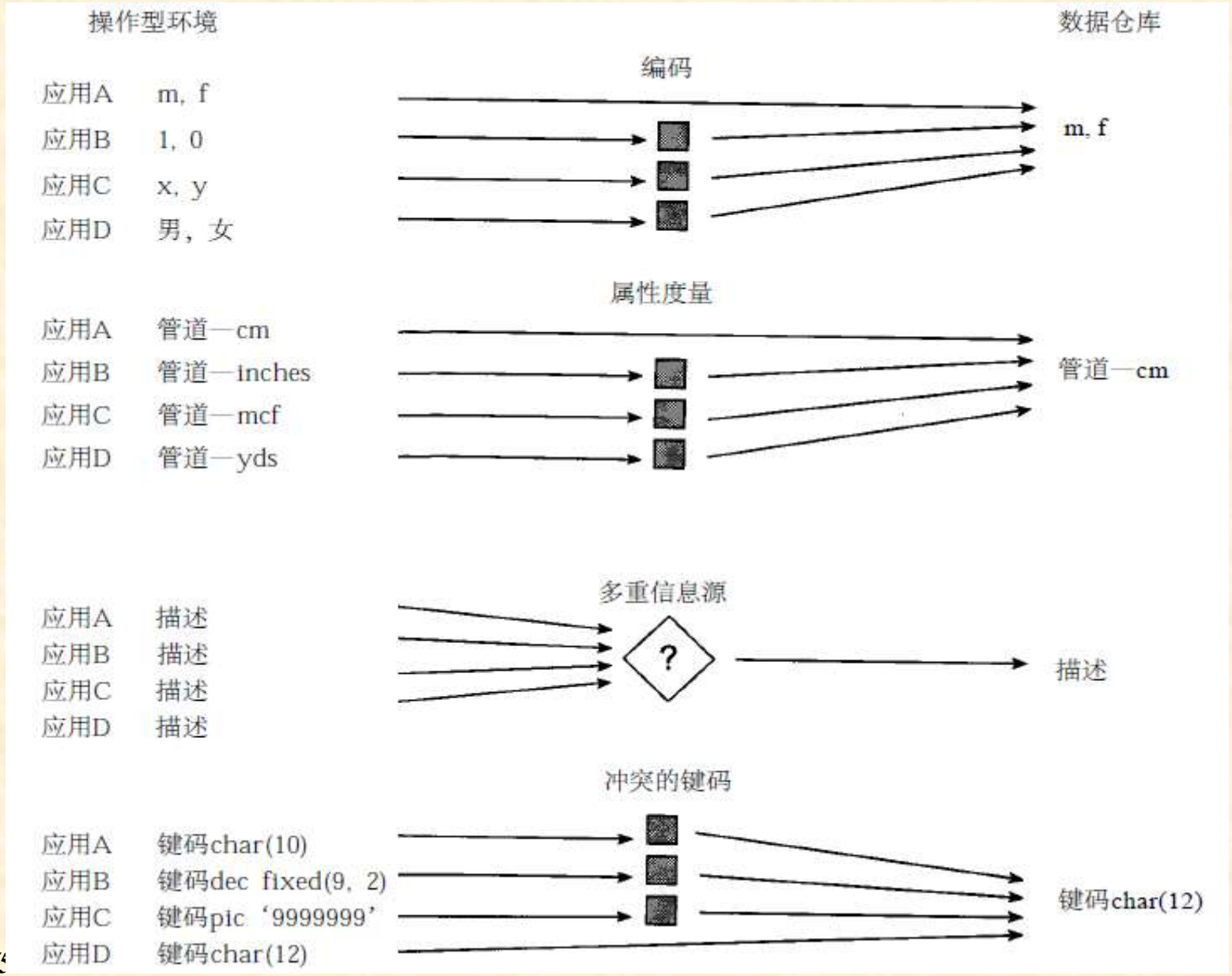
4. 数据仓库的四大特色

（2）集成

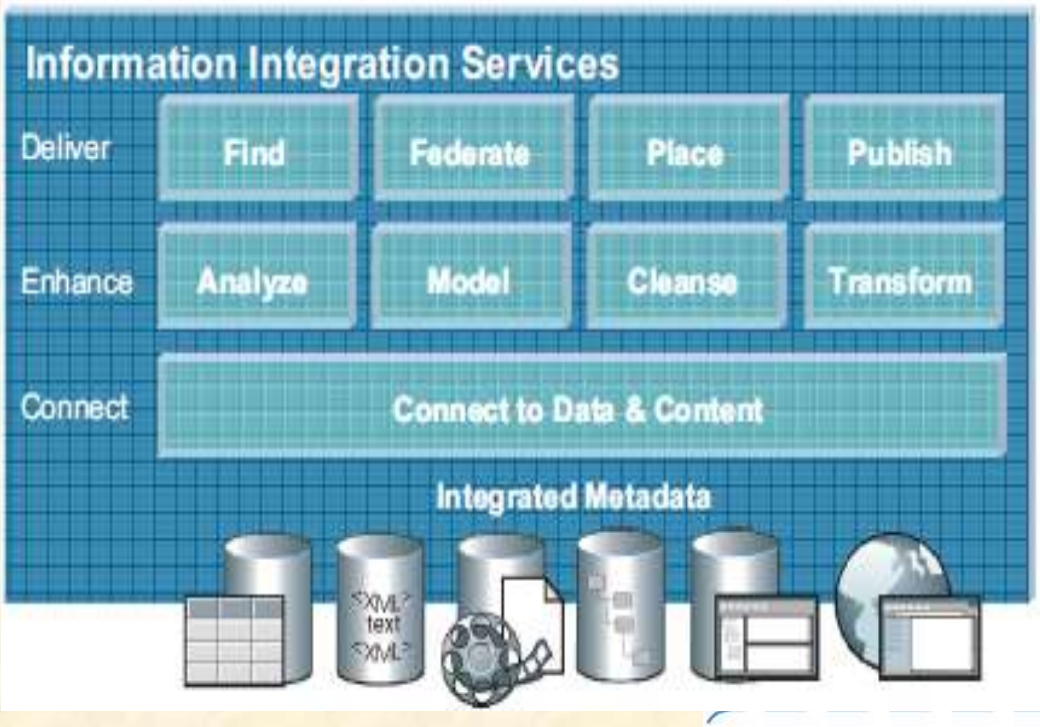
- 数据仓库中的数据是为分析服务的，而分析需要多种广泛的异构数据源以便进行比较、鉴别，因此数据仓库中的数据必须从多个数据源中获取，这些数据源包括多种类型数据库、文件系统以及Internet网上数据等，它们通过数据集成而形成数据仓库中的数据。
- 使用数据清理，数据集成和数据统一技术集成

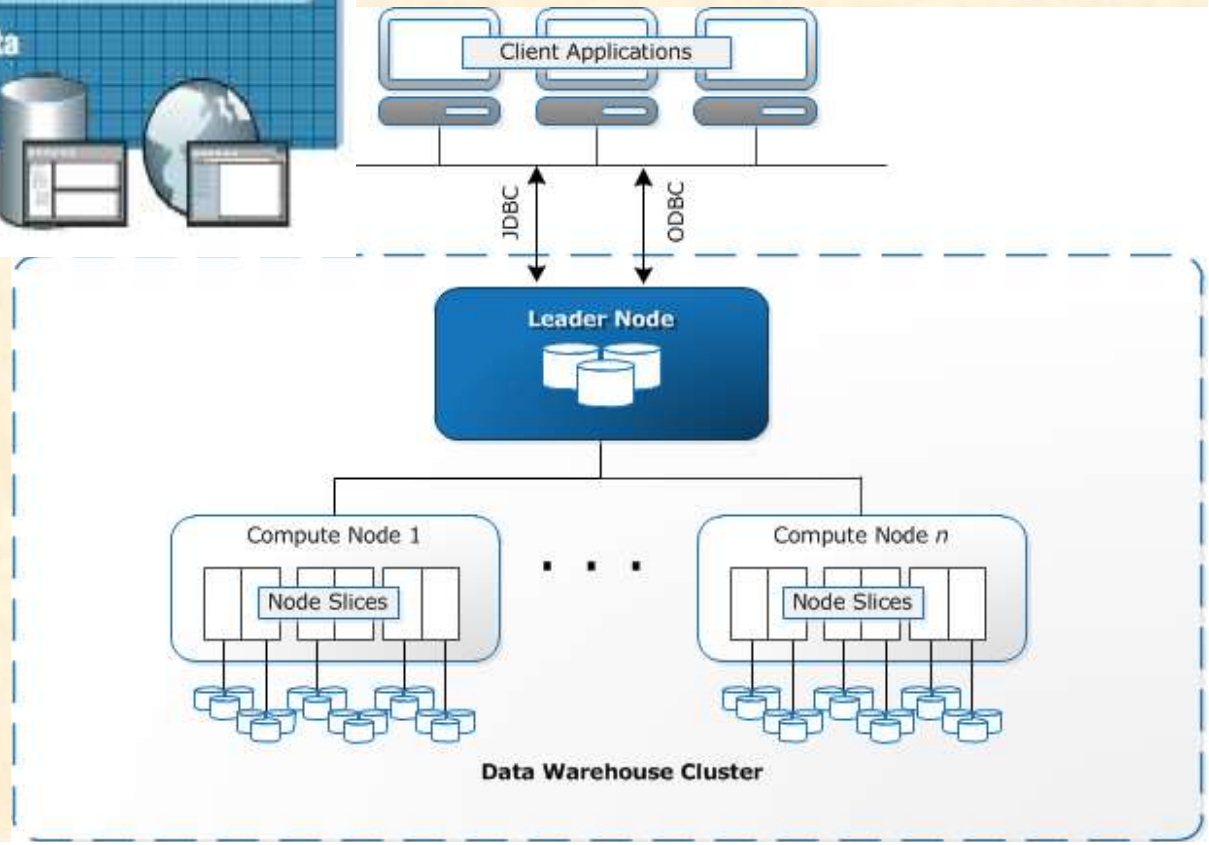
- 集成的方法：
  - 统一
    - 消除不同数据源之间的数据不一致的现象
  - 综合
    - 对原有数据进行综合和计算，如统计、抽样……

4. 数据仓库的四大特色


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)

4. 数据仓库的四大特色

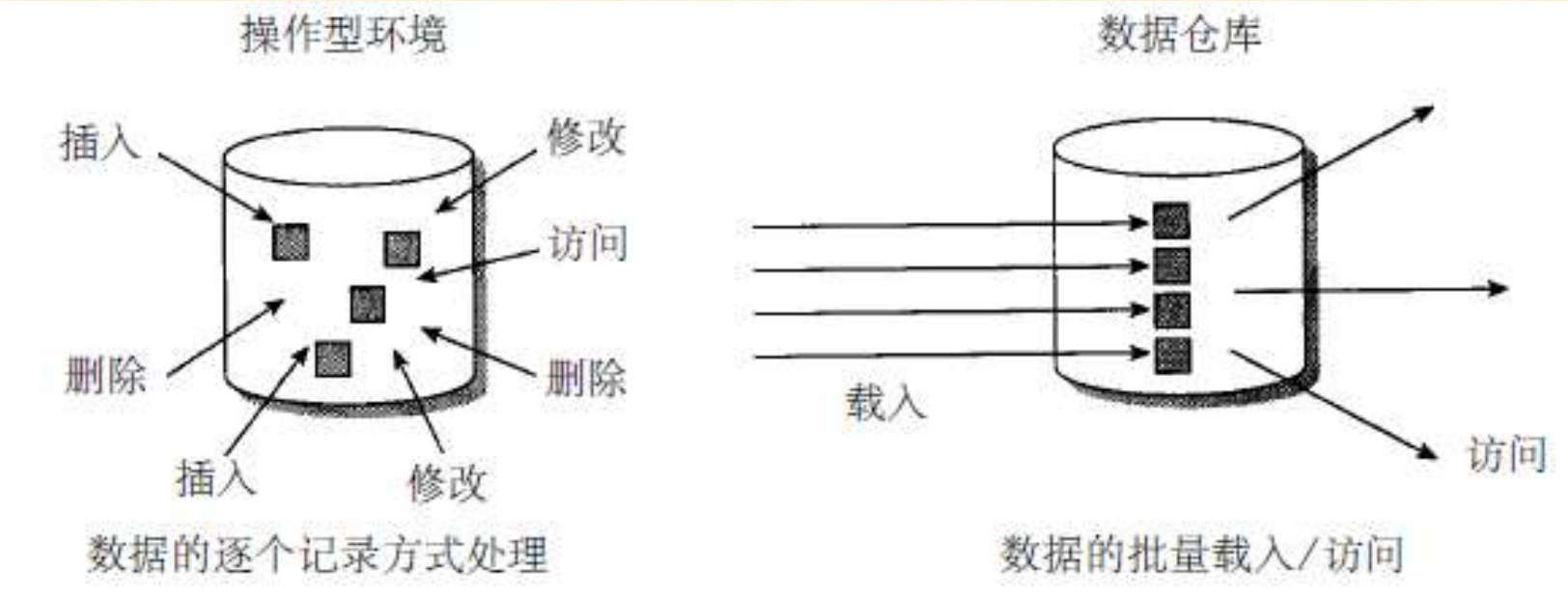

Information Integration Services
Deliver
Find
Federate
Place
Publish
Enhance
Analyze
Model
Cleanse
Transform
Connect
Connect to Data & Content
Integrated Metadata
<XML>
text
</XML>


Client Applications
JDBC
ODBC
Leader Node
Compute Node 1
Node Slices
Compute Node n
Node Slices
Data Warehouse Cluster

4. 数据仓库的四大特色

（3）不可更新

– 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，物理地分离存放数据，执行的主要是‘初始化装入’、‘查询’操作，一般情况下不执行‘更新’操作，不需要事务处理，恢复和并发控制机制。同时，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。



4. 数据仓库的四大特色

（3）不可更新

- 不进行一般意义上的数据更新
- 但也不等于数据仓库中的数据不需要‘更新’操作，如：
  - 在需要进行新的分析决策时，可能需要进行新的数据抽取（‘插入’操作）和‘更新’操作
  - 数据仓库中的一些超过规定的存储期限的数据，也可以通过‘删除’操作丢弃掉
- 因此, 数据仓库的存储管理相对于**DBMS**来说要简单得多

4. 数据仓库的四大特色

（4）随时间不断变化

- 数据仓库内的信息并不只是关于企业当时或某一时点的信息，而是系统记录了企业从过去某一时点到目前的各个阶段的信息，通过这些信息可以对企业的发展历程和未来趋势作出定量分析和预测。

- 因此数据仓库中的数据通常都带有时间属性，同时必须以一定时间段为单位进行统一更新。如：

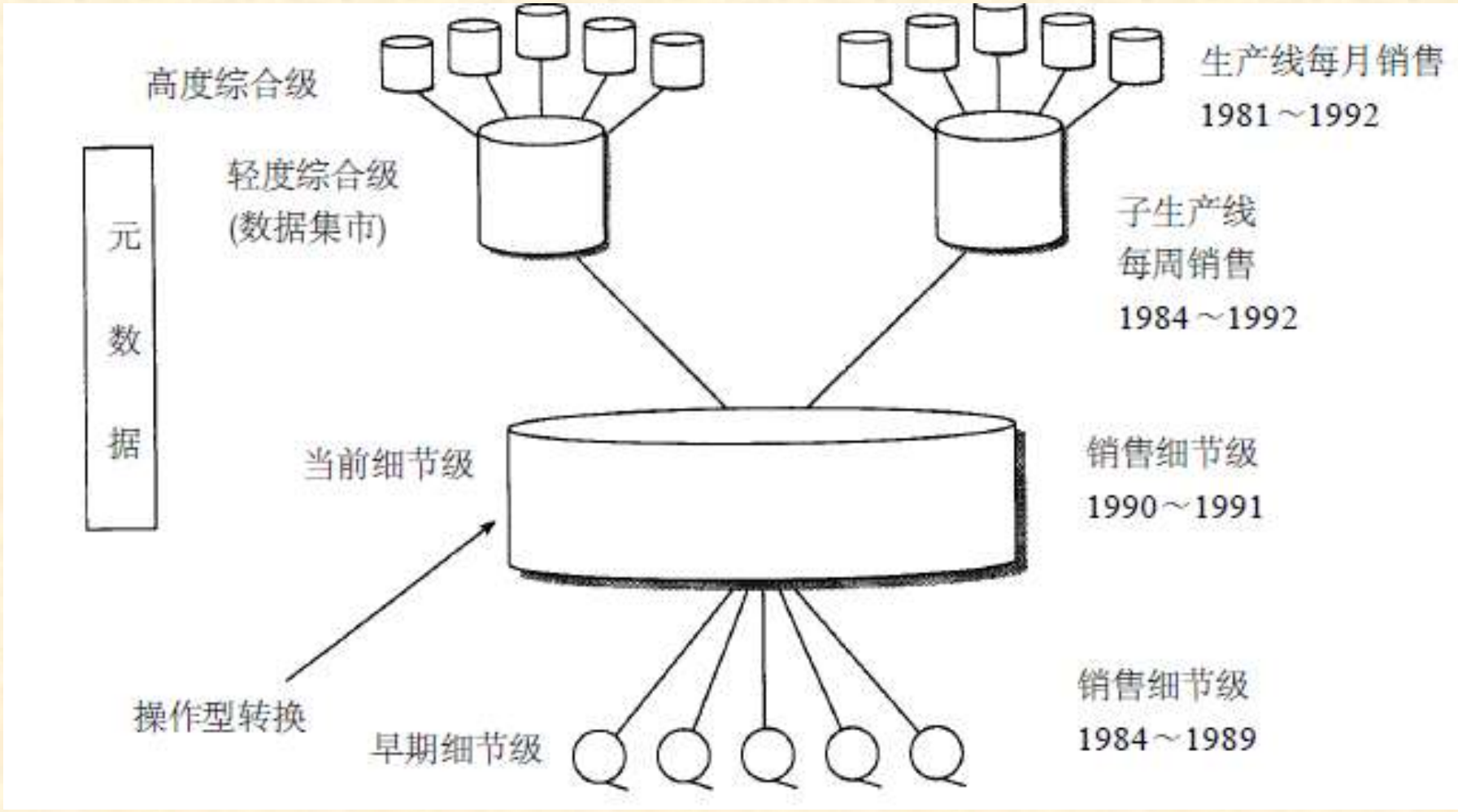
  - 不断增加新的数据内容
  - 不断删去旧的数据内容
  - 更新与时间有关的综合数据

4. 数据仓库的四大特色

（4）随时间不断变化

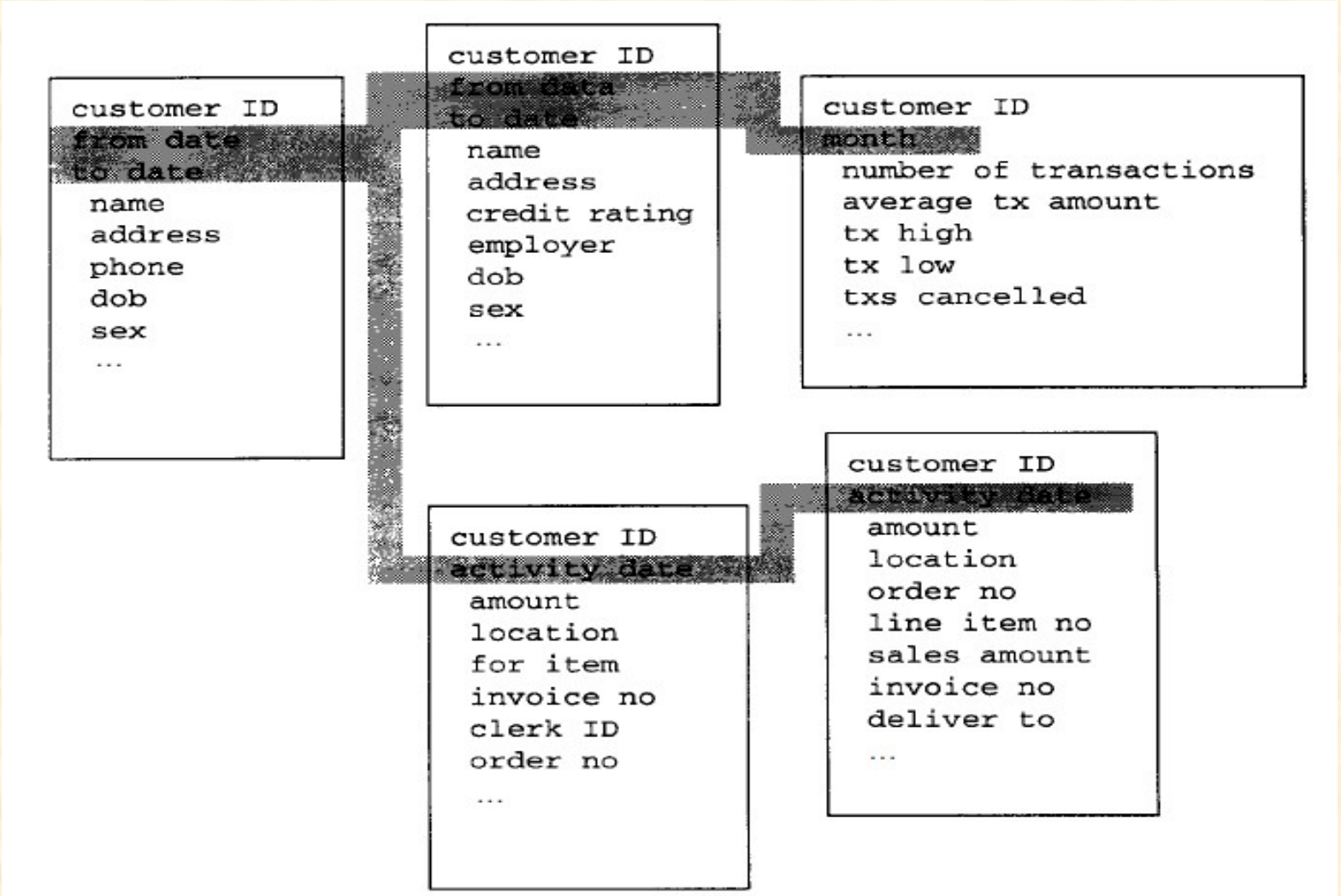
- 数据仓库中的数据时间期限要远远长于数据库系统中的数据时间期限。
  - 操作型数据库系统的时间期限一般是60～90天，而数据仓库中数据的时间期限通常是5～10年。
  - 操作型数据库含有“当前值”的数据，这些数据的准确性在访问时是有效的，同样当前值的数据能被更新。
  - 数据仓库中的数据仅仅是一系列某一时刻生成的复杂的快照。

4. 数据仓库的四大特色



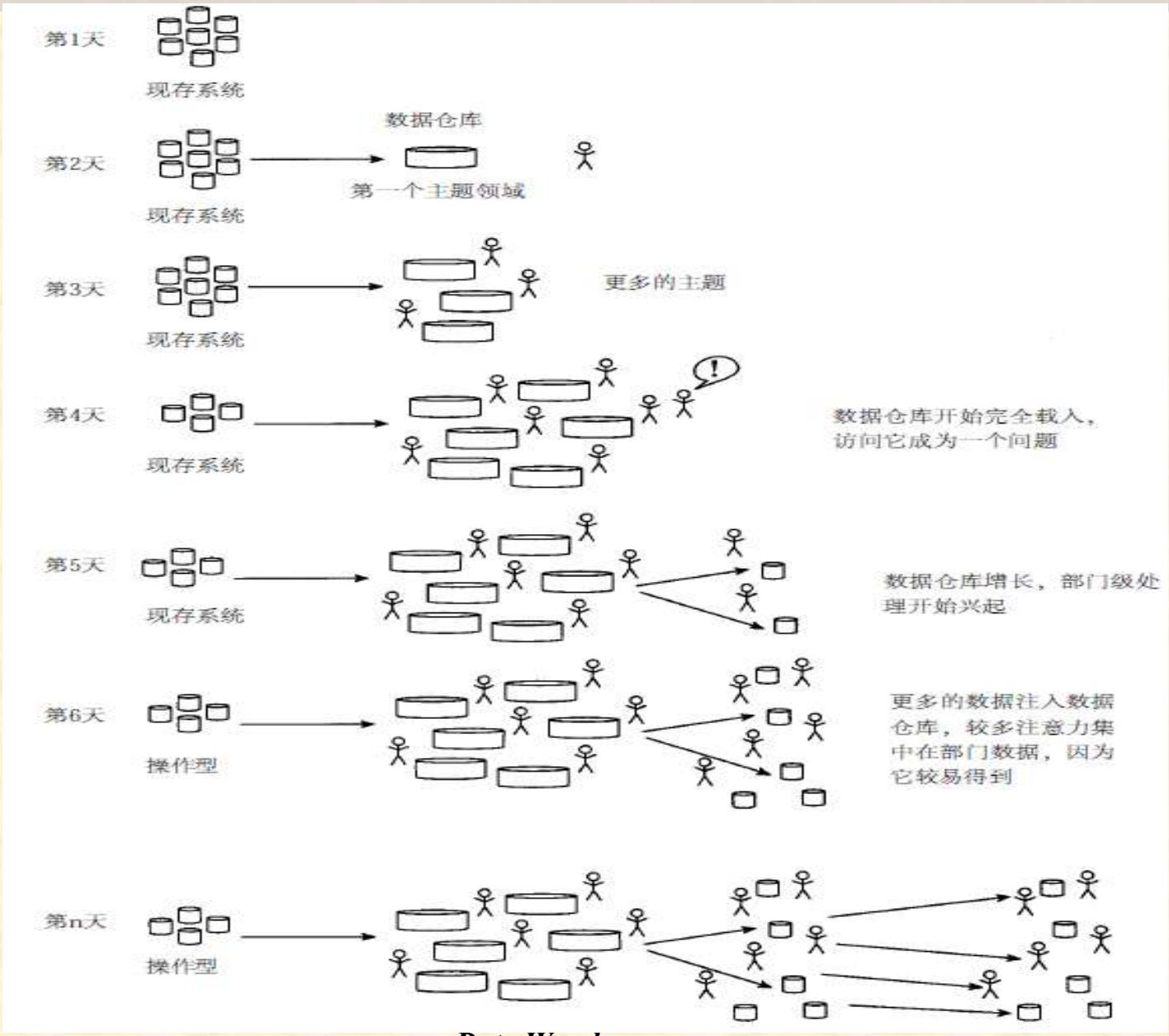
数据仓库中的数据是一系列
某一时刻生成的复杂的数据快照。

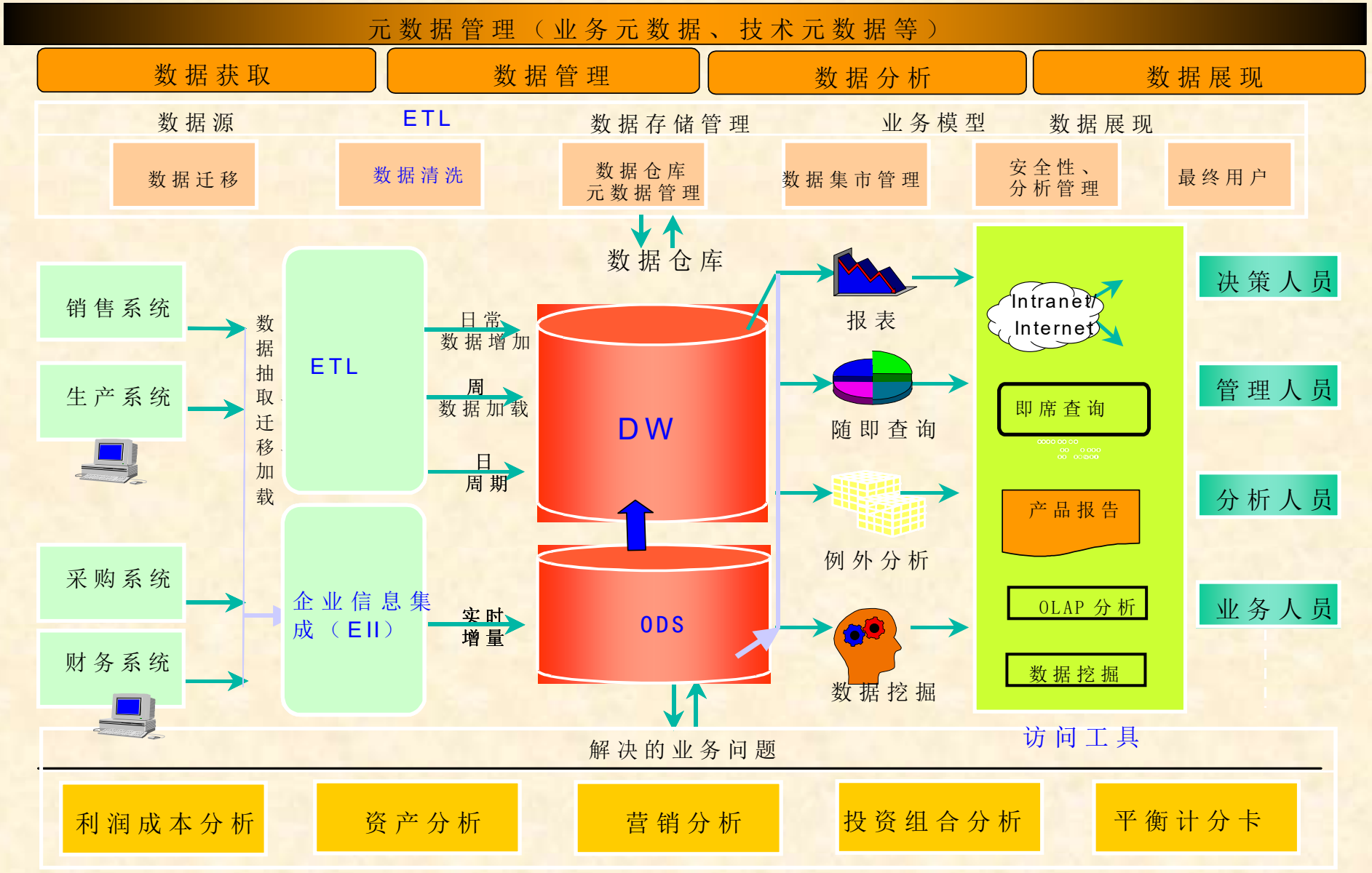
4. 数据仓库的四大特色

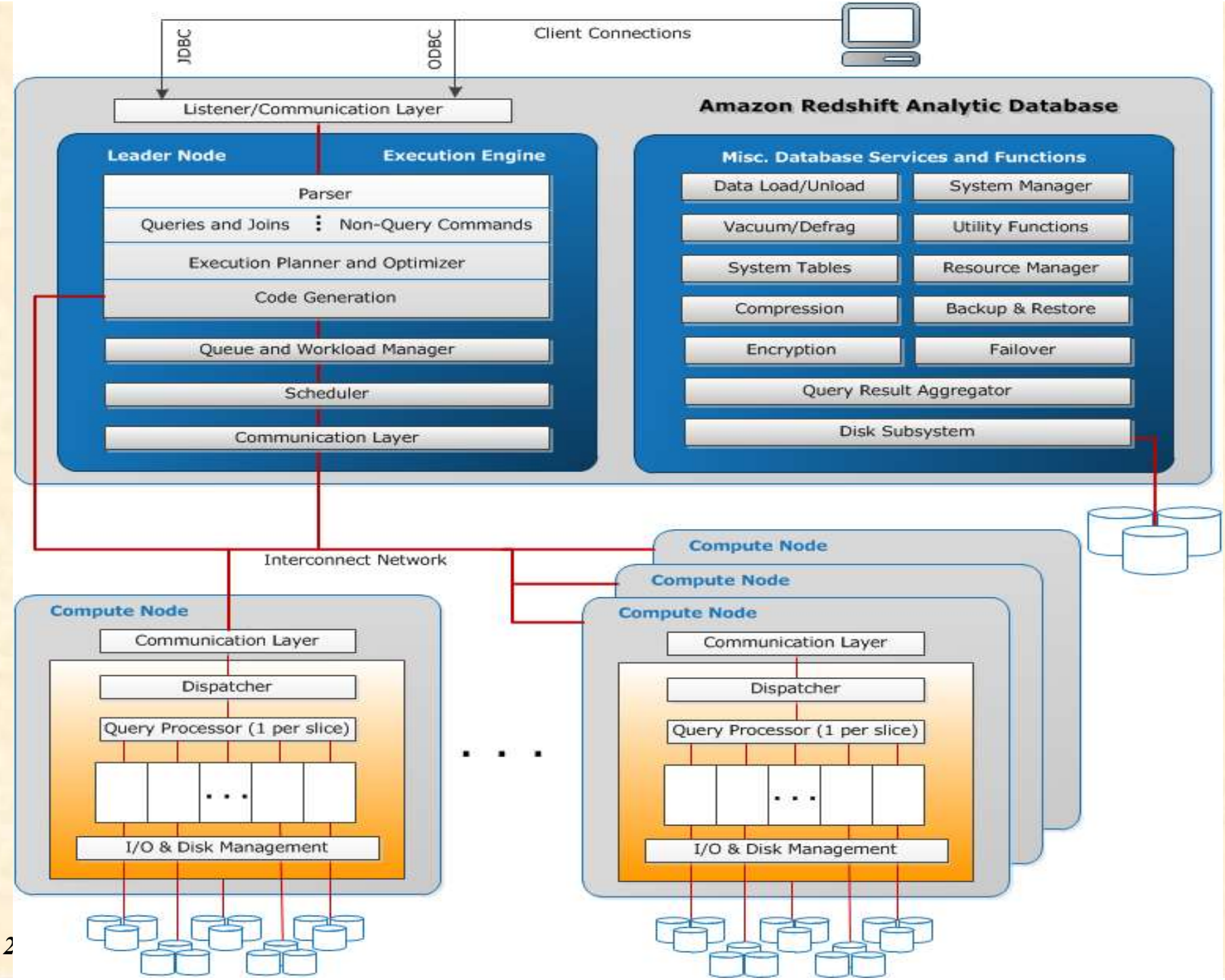


操作型数据的键码可能包含，也可能不包含时间元素,而数据仓库的键码总是包含时间元素。

数据仓库的进化时间图


第1天
现存系统
数据仓库
第2天
现存系统
第一个主题领域
第3天
现存系统
更多的主题
第4天
现存系统
!
数据仓库开始完全载入，
访问它成为一个问题
第5天
现存系统
数据仓库增长，部门级处
理开始兴起
第6天
操作型
更多的数据注入数据
仓库，较多注意力集
中在部门数据，因为
它较易得到
第n天
操作型


元数据管理（业务元数据、技术元数据等）
数据获取
数据管理
数据分析
数据展现
数据源
ETL
数据存储管理
业务模型
数据展现
数据迁移
数据清洗
数据仓库
元数据管理
数据集市管理
安全性、
分析管理
最终用户
数据仓库
销售系统
生产系统
采购系统
财务系统
数据抽取迁移加载
ETL
企业信息集成（EII）
日常
数据增加
周
数据加载
日
周期
实时
增量
DW
ODS
报表
随即查询
例外分析
数据挖掘
Intranet/
Internet
即席查询
产品报告
OLAP分析
数据挖掘
访问工具
决策人员
管理人员
分析人员
业务人员
解决的业务问题
利润成本分析
资产分析
营销分析
投资组合分析
平衡计分卡


JDBC
ODBC
Client Connections
Listener/Communication Layer
Amazon Redshift Analytic Database
Leader Node
Execution Engine
Parser
Queries and Joins
Non-Query Commands
Execution Planner and Optimizer
Code Generation
Queue and Workload Manager
Scheduler
Communication Layer
Misc. Database Services and Functions
Data Load/Unload
System Manager
Vacuum/Defrag
Utility Functions
System Tables
Resource Manager
Compression
Backup & Restore
Encryption
Failover
Query Result Aggregator
Disk Subsystem
Interconnect Network
Compute Node
Compute Node
Compute Node
Communication Layer
Dispatcher
Query Processor (1 per slice)
I/O & Disk Management
Compute Node
Communication Layer
Dispatcher
Query Processor (1 per slice)
I/O & Disk Management