商务智能
Business Intelligence

数据采集与集成

朱卫平 博士
计算机学院
武汉大学

目录

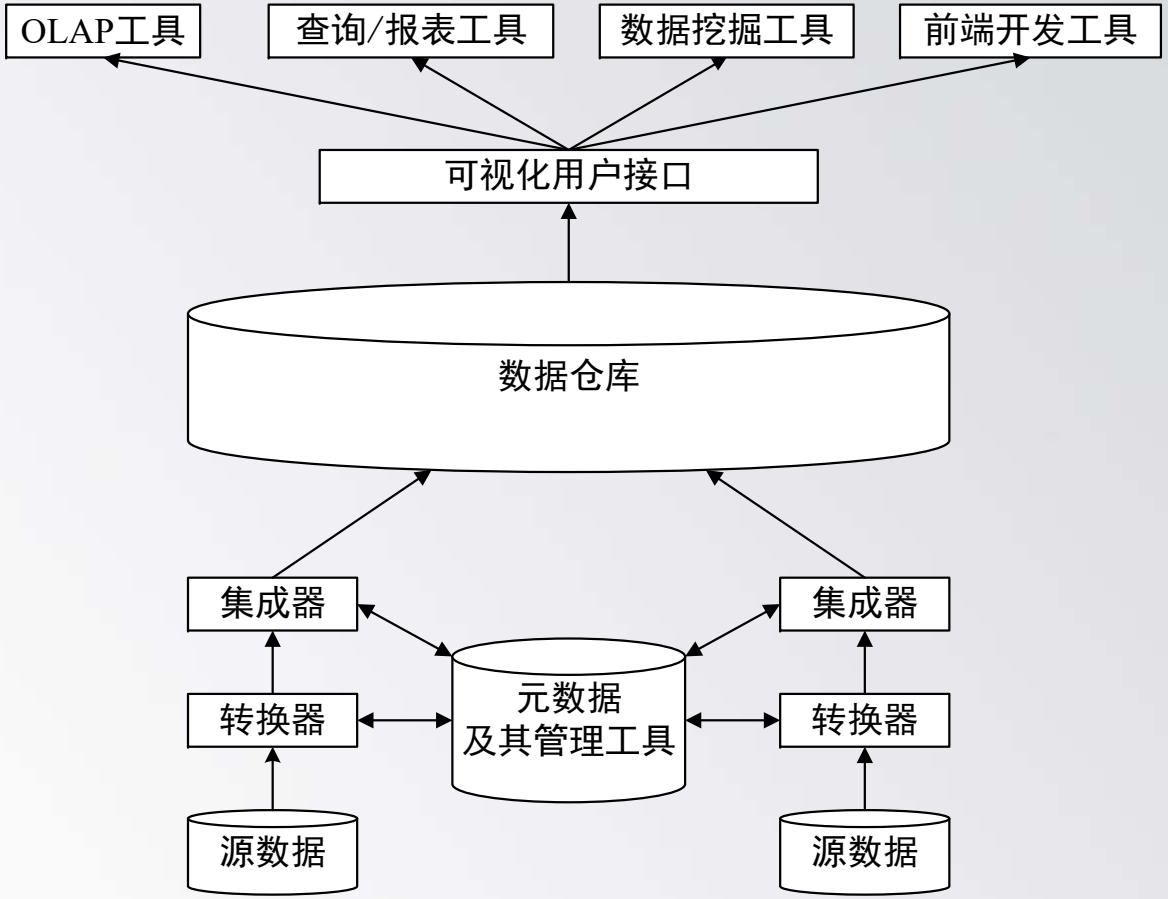
- 数据采集
- 数据集成

数据仓库

- 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。
- 以1992年数据仓库之父W. H. Inmon出版《Building the Data Warehouse》为标志，数据仓库发展速度很快。

数据仓库架构

- 数据仓库从数据源经过转换、集成后获得，同时提供给可视化用户接口用于上层分析。



数据仓库

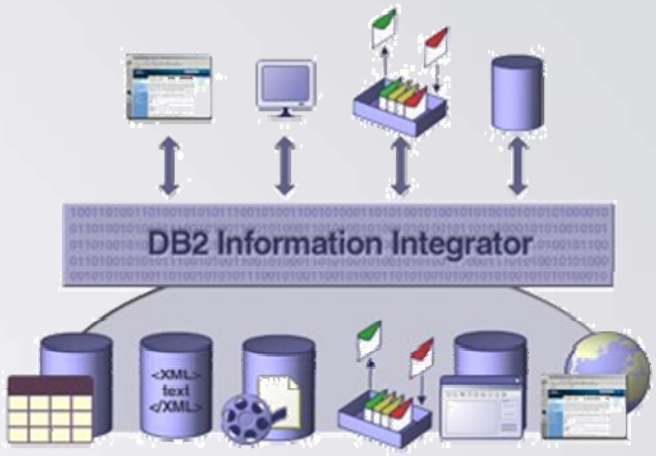
- 数据仓库是与操作型系统相分离的。
- W H Inmon对数据仓库所下的定义：数据仓库是面向主题的、集成的、稳定的、随时间变化的数据集合，用以支持管理决策的过程。
  - 面向主题
  - 集成性
  - 数据的非易失性
  - 数据的时变性

数据仓库特性：面向主题

- 数据仓库中的数据是按照各种主题来组织的。主题在数据仓库中的物理实现是一系列的相关表，这不同于面向应用环境
  - 如保险公司按照应用组织可能是汽车保险、生命保险、伤亡保险，而数据仓库是按照客户、政策、保险金和索赔来组织数据。
  - 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的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。
  - 主题相关的数据通常分布在多个系统中。

数据仓库特性：集成性

- 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。



数据仓库特性：数据的非易失性

- 数据仓库中的数据反映的是一段时间内历史数据的内容，是不同时点的数据库快照的集合，以及基于快照进行统计、综合和重组的导出数据。
  - 数据仓库内的数据有很长的时间跨度，通常是5-10年。
  - 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。
  - 数据不实时更新，仅定期进行删除和加载

数据仓库特性：数据的时变性

- 数据仓库的数据随时间变化主要变现在以下几个方面：
  - 数据仓库中的数据是按照时间顺序追加的，它们都带有时间属性。
  - 当超过规定的存储期限，数据从仓库中删除，同时定期加载新的数据输入数据仓库
  - 不同类型数据的更新频率是不同的。例如，产品属性的变化每个星期更新一次，地理位置的变化每个月更新一次，销售数据每天更新一次。

课后思考题

- 请自己编写一个爬虫程序获取少量的网络数据
- 请解释为什么需要数据仓库以及建立数据仓库的步骤
- 请解释数据仓库的主要特性

Thank You!

Q&A