Data Warehouse

——Data Warehouse Architecture——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

1

Data Warehouse Architecture

- Review the basic concepts of database
- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation
- From data warehousing to data mining

2

Design of Data Warehouse: A Business Analysis Framework

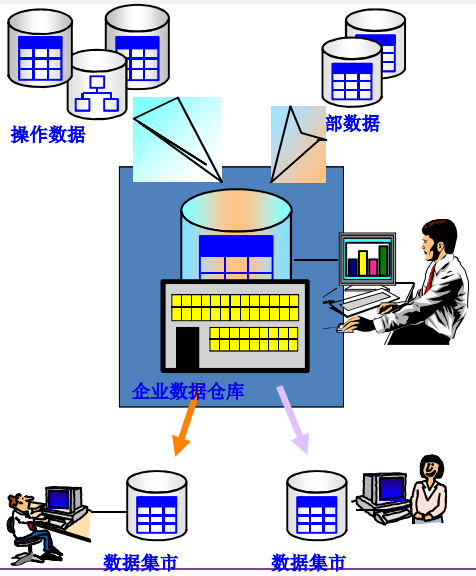
- Four views regarding the design of a data warehouse
  - Top-down view
    - allows selection of the relevant information necessary for the data warehouse
  - Data source view
    - exposes the information being captured, stored, and managed by operational systems
  - Data warehouse view
    - consists of fact tables and dimension tables
  - Business query view
    - sees the perspectives of data in the warehouse from the view of end-user

Data Warehouse Design Process

- Top-down, bottom-up approaches or a combination of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- From software engineering point of view
  - Waterfall(瀑布式): structured and systematic analysis at each step before proceeding to the next
  - Spiral (螺旋式) : rapid generation of increasingly functional systems, short turn around time, quick turn around
- Typical data warehouse design process
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

Data Warehouse — Subject-Oriented

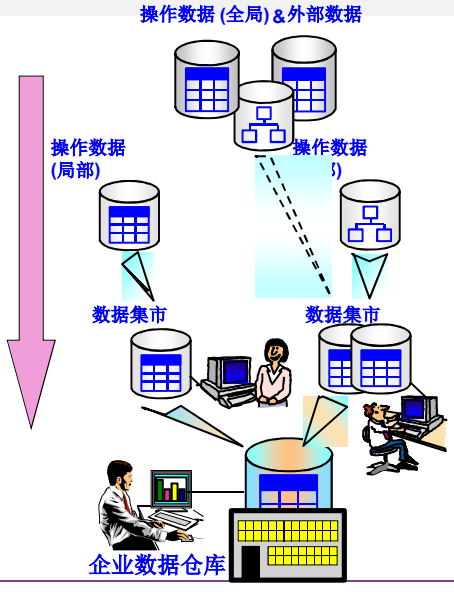
- **自顶向下的数据仓库设计**
- **建造企业数据仓库**
  - 建设中心数据模型
  - 一次性完成数据的重构工作
  - 最小化数据冗余度和不一致性
  - 存储详细的历史数据
- **从企业数据仓库中建造数据集市**
  - 得到大部分的集成数据
  - 直接依赖于数据仓库的可用性
- **问题**
  - 投资效益的时间?
  - 建设中心数据模型的必要性和可能性?
  - 初始费用？



5

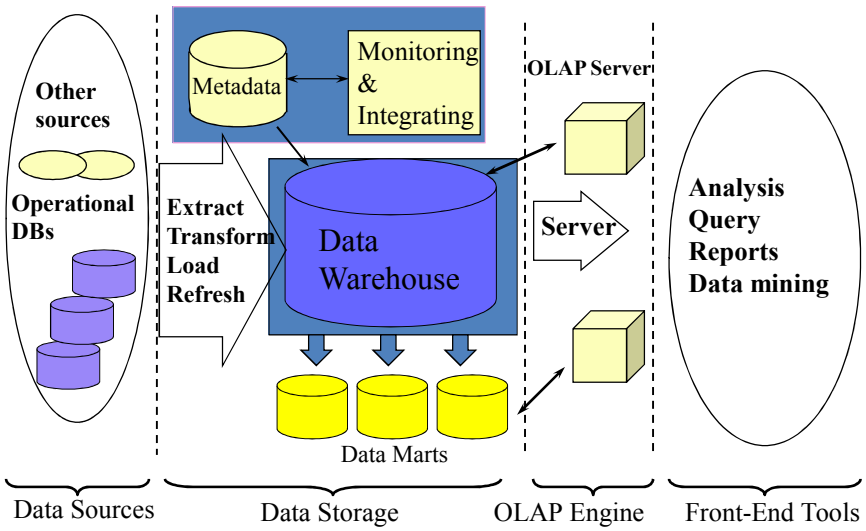
Data Warehouse — Integrated

- **自底向上的数据仓库设计**
- **建立部门数据集市**
  - 限制在一个主题区域
  - 快速投资收益
  - 区域自治 – 设计的可伸缩性强
  - 对相关部门的应用容易复制
  - 对每个数据集市需要数据重构
  - 存在一定的冗余及不一致性
- **逐步扩展到企业数据仓库 (EDW)**
  - 把建造EDW作为一个长期的目标
- **存在的问题:**
  - 数据集市的数据都是可用的吗?
  - 能生成数据模型吗?
  - 如何解决不一致性?



6

Data Warehouse: A Multi-Tiered Architecture



7

附录：关于元数据 —— About Metadata

- **元数据(Meta data): data about data (关于数据的数据)**
- **功能：**
  - 元数据能提供基于用户的信息,如记录数据项的业务描述信息的元数据能帮助用户使用数据。
  - 元数据能支持系统对数据的管理和维护,如关于数据项存储方法的元数据能支持系统以最有效的方式访问数据
- **对于数据仓库的支持：**
  - 描述哪些数据在数据仓库中；
  - 定义要进入数据仓库中的数据和从数据仓库中产生的数据；
  - 记录根据业务事件发生而随之进行的数据抽取工作时间安排；
  - 记录并检测系统数据一致性的要求和执行情况；
  - 衡量数据质量。

8

Thanks !