Data Warehouse

——What is a Data Warehouse?——

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

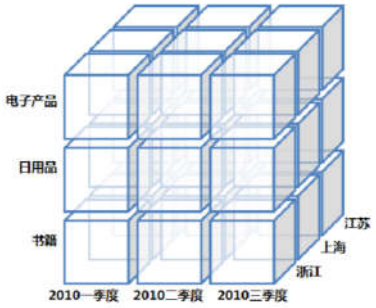
Data Warehouse

- 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

Data Warehouse — Subject-Oriented

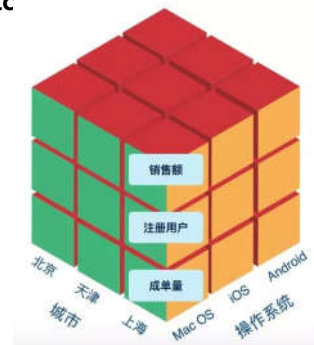
- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.



3

Data Warehouse — Integrated

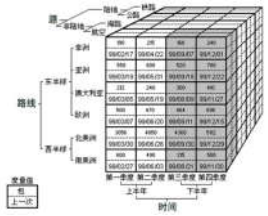
- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc
  - When data is moved to the warehouse, it is converted



4

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element".



5

Data Warehouse—Non-Volatile(非易失的)

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency (并发) control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

6

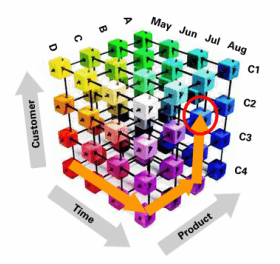
What is a data warehouse?

- **Data warehouse** is a semantically consistent store that serves as a physical implementation of a decision support data model and stores the information on which and enterprise needs to make strategic decisions.
- **Data warehouse** is viewed as an architecture, constructed by integrating data from multiple heterogeneous sources to support structured and/or ad hoc queries, analytical reporting, and decision making.

7

What is data warehouse used for?

- Increasing customer focus
  - buying patterns, buying preference
- Fine-tuning production strategies
  - repositioning(重新配置) products and managing product portfolios (组合) .
- Analyzing operations and looking for sources of profit
- Managing the customer relationships



8

Data Warehouse vs. Heterogeneous DBMS

- Traditional heterogeneous DB integration:
  - Build wrappers/mediators on top of heterogeneous databases
  - Query driven approach
    - When a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set
    - Complex information filtering, compete for resources
- Data warehouse: update-driven, high performance
  - Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

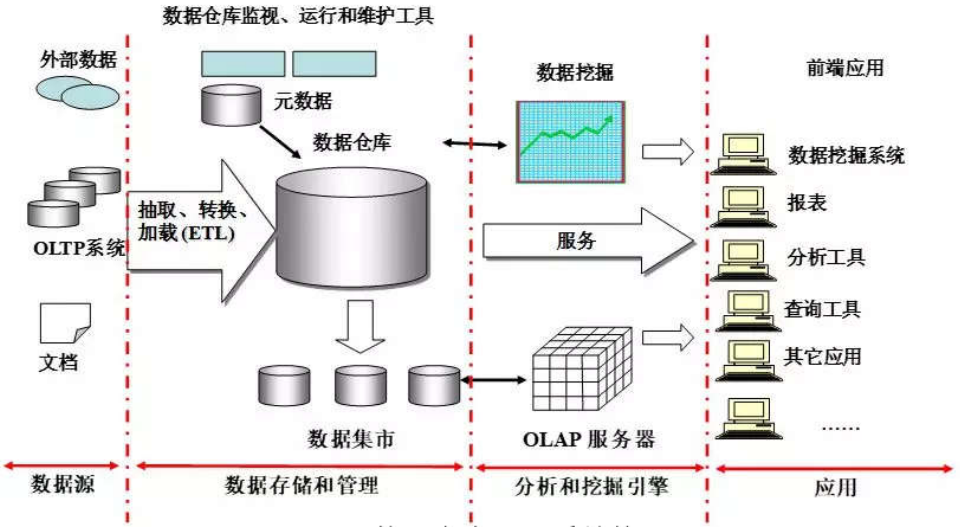
9

Data Warehouse vs. Operational DBMS

- OLTP (On-Line Transaction Processing，联机事务处理)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory (库存) , banking, manufacturing, payroll(工资单), registration, accounting, etc.
- OLAP (On-Line Analytical Processing，联机分析处理)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated (合并统一)
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

10

OLTP vs. OLAP



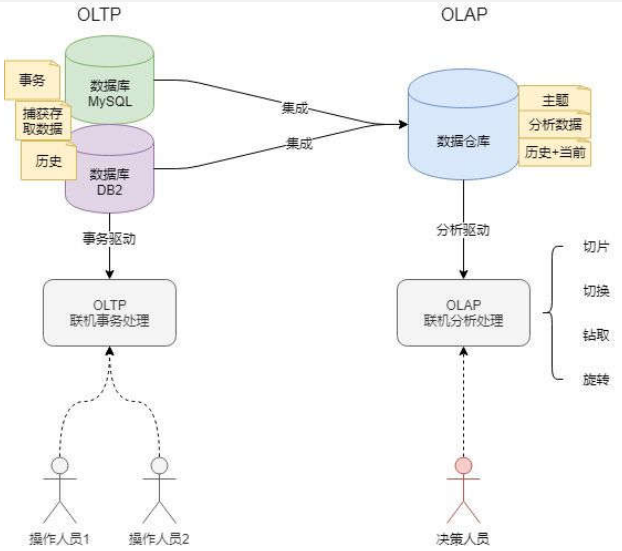
11

OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

12

OLTP vs. OLAP



13

Why Separate Data Warehouse?

- **High performance for both systems**
  - **DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery**
  - **Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation**
- **Different functions and different data:**
  - **missing data(缺失): Decision support requires historical data which operational DBs do not typically maintain**
  - **data consolidation (整合): DS requires consolidation (aggregation, summarization) of data from heterogeneous sources**
  - **data quality (质量): different sources typically use inconsistent data representations, codes and formats which have to be reconciled (一致化处理)**
- **Note: There are more and more systems which perform OLAP analysis directly on relational databases**

14

Thanks !