Data Warehousing

Learning Outcomes

- Define data warehouse
- Difference between data warehouse & DB
- Concepts OLAP & OLTP
- Why separate data warehouse ?

What is Data Warehouse?

- Defined in many different ways, but not rigorously
  - A decision support database that is maintained separately from the organisation's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis
- Definition by Inmon
  - "A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process"
- Data warehousing
  - The process of constructing and using data warehouses

Data Warehouse—Subject-Oriented

- Organised around major subjects, such as customer, product, sales
- Focusing on the modelling 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

Data Warehouse—Integrated

- Integrate multiple heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Apply techniques of Data cleaning and data integration
  - 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

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”

Data Warehouse—Non-Volatile

- Physically separate stores 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 control mechanisms
  - Requires only two operations in data accessing:

    initial loading of data and *access of data*

Data Warehouse vs. Heterogeneous DB

- 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 access and analysis

Data Warehouse vs. Operational DB

- OLTP (On-Line Transaction Processing)
  - Major task of traditional relational DB
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (On-Line Analytical Processing)
  - 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
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

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

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: Decision support requires historical data which operational DBs do not typically maintain
  - Data consolidation: DS requires consolidation (aggregation, summarisation) of data from heterogeneous sources
  - Data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled