Motivation

• Increasingly, organizations are analyzing current and historical data to identify useful patterns and support business strategies.

• Emphasis is on complex, interactive, exploratory analysis of very large datasets created by integrating data from across all parts of an enterprise; data is fairly static.

  – Contrast such **On-Line Analytic Processing (OLAP)** with traditional **On-line Transaction Processing (OLTP)**: mostly long queries, instead of short update Xacts.
What is a Data Warehouse?

• Defined in many different ways, but not rigorously.
  – A decision support database that is maintained separately from the organization’s operational database
  – Support information processing by providing a solid platform of consolidated, historical data for analysis.

• “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.” —W. H. Inmon

• Data warehousing:
  – The process of constructing and using data warehouses
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.
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.
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”.
Non-Volatile

• 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 control mechanisms
  
  – Requires only two operations in data accessing:
    
    • **initial loading of data and access of data**.
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
Data Warehouse vs. Operational DBMS

• **OLTP (on-line transaction processing)**
  – Major task of traditional relational DBMS
  – 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

<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>users</strong></td>
<td>clerk, IT professional</td>
<td>knowledge worker</td>
</tr>
<tr>
<td><strong>function</strong></td>
<td>day to day operations</td>
<td>decision support</td>
</tr>
<tr>
<td><strong>DB design</strong></td>
<td>application-oriented</td>
<td>subject-oriented</td>
</tr>
<tr>
<td><strong>data</strong></td>
<td>current, up-to-date detailed, flat relational isolated</td>
<td>historical, summarized, multidimensional integrated, consolidated</td>
</tr>
<tr>
<td><strong>usage</strong></td>
<td>repetitive</td>
<td>ad-hoc</td>
</tr>
<tr>
<td><strong>access</strong></td>
<td>read/write</td>
<td>lots of scans</td>
</tr>
<tr>
<td></td>
<td>index/hash on prim. key</td>
<td></td>
</tr>
<tr>
<td><strong>unit of work</strong></td>
<td>short, simple transaction</td>
<td>complex query</td>
</tr>
<tr>
<td><strong># records accessed</strong></td>
<td>tens</td>
<td>millions</td>
</tr>
<tr>
<td><strong>#users</strong></td>
<td>thousands</td>
<td>hundreds</td>
</tr>
<tr>
<td><strong>DB size</strong></td>
<td>100MB-GB</td>
<td>100GB-TB</td>
</tr>
<tr>
<td><strong>metric</strong></td>
<td>transaction throughput</td>
<td>query throughput, response</td>
</tr>
</tbody>
</table>
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, summarization) of data from heterogeneous sources
  – **data quality**: different sources typically use inconsistent data representations, codes and formats which have to be reconciled