DATA WAREHOUSE

WHAT IS DATA WAREHOUSE?
“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 as relational databases, flat files, on-line transaction records
Data cleaning and data integration techniques are applied To 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
a) Operational database: current value data
b) Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
Every key structure in the data warehouse
a) Contains an element of time, explicitly or implicitly
b) But the key of operational data may or may not contain “time element”

NONVOLATILE
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

Author: Mohammed Abdul Khaliq
Email: dwikatmo@hotmail.com
Topic: Data Mining / Data Warehouse
Date: 09/03/2011
Page: 1 of 3
DATA MART
A data mart is a repository of data gathered from operational data and other sources that is designed to serve a particular community of knowledge workers

OLTP (ON-LINE TRANSACTION PROCESSING)
Provides you with a very good view of what is happening, but cannot predict what will happen in the future or why it is happening
✓ Major task of traditional relational DBMS
✓ Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

OLAP (ON-LINE ANALYTICAL PROCESSING)
A database technology that has been specially designed to deal with high performance querying and reporting. OLAP data is organized hierarchically and stored in cubes instead of tables.
✓ Major task of data warehouse system
✓ Data analysis and decision making

OLTP VERSUS OLAP

<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>users</td>
<td>clerk, IT professional</td>
<td>knowledge worker</td>
</tr>
<tr>
<td>function</td>
<td>day to day operations</td>
<td>decision support</td>
</tr>
<tr>
<td>DB design</td>
<td>application-oriented</td>
<td>subject-oriented</td>
</tr>
<tr>
<td>usage</td>
<td>repetitive</td>
<td>ad-hoc</td>
</tr>
<tr>
<td>access</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>data</td>
<td>current, up-to-date</td>
<td>historical, summarized, multidimensional</td>
</tr>
<tr>
<td></td>
<td>detailed, flat relational</td>
<td>integrated, consolidated</td>
</tr>
<tr>
<td></td>
<td>isolated</td>
<td></td>
</tr>
<tr>
<td>unit of work</td>
<td>short, simple transaction</td>
<td>complex query</td>
</tr>
<tr>
<td># records accessed</td>
<td>tens</td>
<td>millions</td>
</tr>
<tr>
<td>#users</td>
<td>thousands</td>
<td>hundreds</td>
</tr>
<tr>
<td>DB size</td>
<td>100MB-GB</td>
<td>100GB-TB</td>
</tr>
<tr>
<td>metric</td>
<td>transaction throughput</td>
<td>query throughput, response</td>
</tr>
</tbody>
</table>
## Data Mining Versus Statistical Analysis

<table>
<thead>
<tr>
<th>Statistical Analysis</th>
<th>Data Mining</th>
</tr>
</thead>
</table>
| Hypothesis testing as:  
Is the relationship significant?  
Use a t-test to validate significance | Does not require assumptions to be made about data |
| Tends to rely on sampling | Can find patterns in very large amounts of data |
| Requires strong statistical skills | Requires understanding of data and business problem |
| Tests for statistical correctness of models as:  
Are statistical assumptions of models correct?  
Eg Is the R-Square good? | Originally developed to act as expert systems to solve problems |