Chapter 11: Data Warehousing

Modern Database Management
7th Edition
Jeffrey A. Hoffer, Mary B. Prescott, Fred R. McFadden

© 2005 by Prentice Hall

Definition

- **Data Warehouse:**
  - A subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making processes
  - **Subject-oriented:** e.g. customers, patients, students, products
  - **Integrated:** Consistent naming conventions, formats, encoding structures; from multiple data sources
  - **Time-variant:** Can study trends and changes
  - **Non-updatable:** Read-only, periodically refreshed

- **Data Mart:**
  - A data warehouse that is limited in scope
Need for Data Warehousing

- Integrated, company-wide view of high-quality information (from disparate databases)
- Separation of operational (OLTP) and informational systems and data (for improved performance)

### Table 11-1 Comparison of Operational and Informational Systems

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Operational Systems</th>
<th>Informational Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary purpose</td>
<td>Run the business on a current basis</td>
<td>Support managerial decision making</td>
</tr>
<tr>
<td>Type of data</td>
<td>Current representation of state of the business</td>
<td>Historical point-in-time (snapshots) and predictions</td>
</tr>
<tr>
<td>Primary users</td>
<td>Clerks, salespersons, administrators</td>
<td>Managers, business analysts, customers</td>
</tr>
<tr>
<td>Scope of usage</td>
<td>Narrow, planned, and simple updates and queries</td>
<td>Broad, ad hoc, complex queries and analysis</td>
</tr>
<tr>
<td>Design goal</td>
<td>Performance throughput, availability</td>
<td>Ease of flexible access and use</td>
</tr>
<tr>
<td>Volume</td>
<td>Many, constant updates and queries on one or a few table rows</td>
<td>Periodic batch updates and queries requiring many or all rows</td>
</tr>
</tbody>
</table>

Data Reconciliation

- Typical operational data is:
  - Transient – not historical
  - Not normalized (perhaps due to denormalization for performance)
  - Restricted in scope – not comprehensive
  - Sometimes poor quality – inconsistencies and errors

- After ETL (Extract-Transform-Load), data should be:
  - Detailed – not summarized yet
  - Historical – periodic
  - Normalized – 3rd normal form or higher
  - Comprehensive – enterprise-wide perspective
  - Timely – data should be current enough to assist decision-making
  - Quality controlled – accurate with full integrity
Figure 11.13: Components of a star schema

- **Fact tables** contain factual or quantitative data
- **Dimension tables** contain descriptions about the subjects of the business
- 1-many relationship between dimension tables and fact tables

Dimension tables are denormalized to maximize performance.

Data in fact table are called measures (or dependent attributes).

Figure 11.14: Star schema example

**Fact table** provides statistics for sales broken down by product, period and store dimensions.

- **PRODUCT**:
  - Product Code
  - Description
  - Color
  - Size

- **PERIOD**:
  - Period Code
  - Year
  - Quarter
  - Month
  - Day

- **SALES**:
  - Product_Code
  - Period_Code
  - Store_Code
  - Units_Sold
  - Dollars_Sold
  - Dollars_Cost

- **STORE**:
  - Store_Code
  - Store_Name
  - City
  - Telephone
  - Manager
Issues Regarding Star Schema

- Dimension table keys must be surrogate (non-intelligent and non-business related), because:
  - Keys may change over time
  - Length/format consistency
- Granularity of Fact Table – what level of detail do you want?
  - Transactional grain – finest level
  - Aggregated grain – more summarized
  - Finer grains ⇒ better market basket analysis capability
  - Finer grain ⇒ more dimension tables, more rows in fact table
- Duration of the database – how much history should be kept?
  - General guideline: As much as the technology employed allows.
On-Line Analytical Processing (OLAP)

- The use of a set of (graphical) tools that provides users with multidimensional views of their data and allows them to analyze the data using simple windowing techniques

OLAP Operations
- **Dicing** - aggregating "dices" of the cube.
- **Cube slicing** – come up with (2-D) view of part of data based on restricting the dimensions
- **Drill-down** – going from summary to more detailed views

Figure 11-22: Dicing and slicing a data cube
Starting with summary data, users can obtain details for particular cells.

**Figure 11-23: Example of drill-down**

**Summary report**

<table>
<thead>
<tr>
<th>Brand</th>
<th>Package size</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoftTowel</td>
<td>6-pack</td>
<td>$75</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>3-pack</td>
<td>$100</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>6-pack</td>
<td>$50</td>
</tr>
</tbody>
</table>

**Drill-down with color added**

<table>
<thead>
<tr>
<th>Brand</th>
<th>Package size</th>
<th>Color</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoftTowel</td>
<td>2-pack</td>
<td>White</td>
<td>$30</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>2-pack</td>
<td>Yellow</td>
<td>$25</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>2-pack</td>
<td>Pink</td>
<td>$20</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>6-pack</td>
<td>White</td>
<td>$50</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>6-pack</td>
<td>Green</td>
<td>$25</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>3-pack</td>
<td>Yellow</td>
<td>$25</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>6-pack</td>
<td>White</td>
<td>$30</td>
</tr>
<tr>
<td>SoftTowel</td>
<td>6-pack</td>
<td>Yellow</td>
<td>$20</td>
</tr>
</tbody>
</table>