Core Concepts in DW

- From database to data warehouse
- Data warehouse architecture
- Data organization in data warehouse
From database to data warehouse

The Spider’s Web problem

- Efficient access to information is key to business decision and success
- Different departments require different types of information
  - Sales department is interested in sales and marketing information, not research or production information
  - Need to extract specific data from cooperate database for the department database
From data base to data warehouse

The Spider’s Web problem (con’t)

- Different users in a department require different categories of information
  - Need to extract data from various department databases for personal database

- The inter-connection between personal databases, department databases and cooperate database forms a complex Spider Web.

- The bigger the cooperate, the worse the Spider Web problem is.
From data base to data warehouse

Here are some problems it may cause

- Unreliable data analysis results
- Inefficient data processing
- Difficult to convert data to useful information
From data base to data warehouse

- Unreliable data analysis results

- Proposal 1: Market UP
- Proposal 2: Market DOWN
From data base to data warehouse

- Inefficient data processing
  - To create a cooperate operational report for the CEO may take lots of man power, technology, and time to extract information from Spider Web
  - Unless report requirements are known before hand, similar effort must be spent for every report every time
From data base to data warehouse

- Difficult to convert data to useful information
  - A telecom company tries to analyze a customer’s sales of this year compared to the past 3 years.
  - It requires the usage data, fees data, payment data, customer service data, …
  - The data extracted from the Spider Web may not provide all information
    - E.g. different formats, different time frames, …
From data base to data warehouse

### OLTP vs OLAP

<table>
<thead>
<tr>
<th>OLTP systems</th>
<th>Data warehousing systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holds current data</td>
<td>Holds historical data</td>
</tr>
<tr>
<td>Stores detailed data</td>
<td>Stores detailed, lightly, and highly summarized data</td>
</tr>
<tr>
<td>Data is dynamic</td>
<td>Data is largely static</td>
</tr>
<tr>
<td>Repetitive processing</td>
<td><em>Ad hoc</em>, unstructured, and heuristic processing</td>
</tr>
<tr>
<td>High level of transaction throughput</td>
<td>Medium to low level of transaction throughput</td>
</tr>
<tr>
<td>Predictable pattern of usage</td>
<td>Unpredictable pattern of usage</td>
</tr>
<tr>
<td>Transaction-driven</td>
<td>Analysis driven</td>
</tr>
<tr>
<td>Application-oriented</td>
<td>Subject-oriented</td>
</tr>
<tr>
<td>Supports day-to-day decisions</td>
<td>Supports strategic decisions</td>
</tr>
<tr>
<td>Serves large number of clerical/operational users</td>
<td>Serves relatively low number of managerial users</td>
</tr>
</tbody>
</table>
Data warehouse architecture

- Data Sources
- Data Storage
- OLAP Engine
- Front-End Tools

Data Warehouse

- Metadata
- Extract
- Transform
- Load
- Refresh

- OLAP Engine
- Analysis
- Query
- Reports
- Data mining

- Monitor & Integrator
- Serve

- Other sources
- Operational DBs

- Data Marts
Data warehouse architecture

- Extract/Transform/Load (ETL)
  - Data extract
    - Extract data from databases
    - For customer analysis, extract customer related data, not staff data
  - Data cleaning
    - Customer data may exist under Customer database and Customer Service database using different DBMS
Data warehouse architecture

Data transformation

- Data under different DBMS, e.g. IBM DB2, Oracle, Informix, Sybase, SQL Server may have different formats,
  - Year/month/day
  - Month/day/year
  - Day/month/year
Data warehouse architecture

Data loading

- Load the data from databases according to the physical data model
  - Cleaning data fields, fill-in blanks, check validity

- Many ETL functionalities have been enhanced in current DW tools
Data warehouse architecture

- Data storage (repository)
  - For storing *data warehouse*, *meta data*, and *data marts*

- Data warehouse
  - Contains multidimensional databases, relational databases, integration of the two
Data warehouse architecture

- **Meta data**
  - Contains *management* meta data and *user* meta data

- **Management meta data**
  - For DW designers and tech supports
  - Describes DW design processes and management tasks
Data warehouse architecture

Management meta data

Content includes:

- Data source information
- How data are transformed
- Data types, data location, and data formats in DW
- Rules of data cleaning and insertion
- Data mapping
- Access rights, backup history, archive history, data transmission history, data acquisition history, …
Data warehouse architecture

User meta data

- To assist users query processing, analyzing results, understanding data organization in DW

Content includes:

- Subject areas and information formats, including query, report, chart, audio, and video, …
- Internet web pages
- Other DW information
Data warehouse architecture

- Data mart
  - A smaller data warehouse
  - For one function or one department or one subject
  - Cost less to develop
  - More achievable
Data warehouse architecture

OLAP (On-Line Analytical Processing)

- Provide decision support information under multi-dimension environment with specific queries and reports.
- Multi-dimension: analyzing sales under time, location, and product dimensions
- Multi-dimensional analysis
  - Operations such as slice, dice, drill-down, roll-up, pivot, …
- Multi-dimensional model
  - ROLAP, MOLAP, hybrid, …
Data organization in data warehouse

- Storage structure in DW
- Data granularity
- Data organization
- Data insertion and deletion
Data structures in DW

Storage structure in DW
Data structures in DW

Storage structure in DW

- Four levels: highly summarized, lightly summarized, current detail, old detail
- Different data grain
- High volume lower level data may get “old” and be removed after for example 15 years
- Low volume in higher level data, usually kept in system for longer time
Data structures in DW

Data granularity

Data grain

- The level of data integration or aggregation
- The higher in the hierarchy, the coarser is the grain
  - Sales by day, by week, by month, by year
  - Sales by city, by state, by region, by country, by continent

Sampling

- Data in DW are sampled from detailed data
Data structures in DW

- Data granularity
  - Data partition
    - Integrate or aggregate according to independent categories or periods

<table>
<thead>
<tr>
<th></th>
<th>Health insurance</th>
<th>Life insurance</th>
<th>Car insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Slice 1</td>
<td>Slice 2</td>
<td>Slice 3</td>
</tr>
<tr>
<td>2006</td>
<td>Slice 4</td>
<td>Slice 5</td>
<td>Slice 6</td>
</tr>
<tr>
<td>2007</td>
<td>Slice 7</td>
<td>Slice 8</td>
<td>Slice 9</td>
</tr>
</tbody>
</table>
Data structures in DW

- Data organization
  - Simple data
    - E.g., day by day data in the DW
  - Fixed data
    - E.g., weekly, monthly, quarterly, yearly data in the DW
  - Continuous data
    - Combination of the previous two
### Data structures in DW

#### Data organization

- Continuous data

#### Jan 2007

<table>
<thead>
<tr>
<th>Item #</th>
<th>Description</th>
<th>Location</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pen</td>
<td>New York</td>
<td>2007/1-2007/2</td>
</tr>
<tr>
<td>3</td>
<td>CD</td>
<td>Paris</td>
<td>2007/1</td>
</tr>
<tr>
<td>4</td>
<td>Candy</td>
<td>London</td>
<td>2007/1-2007/2</td>
</tr>
</tbody>
</table>

#### Feb 2007

<table>
<thead>
<tr>
<th>Item #</th>
<th>Description</th>
<th>Location</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pen</td>
<td>New York</td>
<td>2007/1-2007/2</td>
</tr>
<tr>
<td>2</td>
<td>Book</td>
<td>Tokyo</td>
<td>2007/2</td>
</tr>
<tr>
<td>4</td>
<td>Candy</td>
<td>London</td>
<td>2007/1-2007/2</td>
</tr>
</tbody>
</table>
Data structures in DW

- Data insertion and deletion
- New data to be added
  - Time index method
  - Snap-shot comparison method
  - Delta file method
  - Log file method
- Data deletion
  - Not just deleting detailed data
  - Remove all granularity of data
Home work

- What are the four characteristics of data in a data warehouse?
- What are the processes to convert data from transaction database to data warehouse?
- What are data granularity, data partition, data insertion, and data deletion?