Data Warehouse Design Models

- Design overview
- Three models
  - Conceptual model
  - Logical model
  - Physical model
- Dimensionality modeling
Database design and data warehouse design are different

<table>
<thead>
<tr>
<th></th>
<th>Database design</th>
<th>DW design</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processing type</strong></td>
<td>Subject application</td>
<td>Subject analysis</td>
</tr>
<tr>
<td><strong>Application requirement</strong></td>
<td>Clear, specific</td>
<td>Not clear</td>
</tr>
<tr>
<td><strong>Design goal</strong></td>
<td>Open, secure, high performance</td>
<td>Four characteristics of DW and consistency</td>
</tr>
<tr>
<td><strong>Design methodology</strong></td>
<td>Requirement-driven</td>
<td>data-driven</td>
</tr>
</tbody>
</table>
Three models

- Conceptual model, Logical model, Physical model, or
- High-level model, middle-level model, low-level model (Inmon)
  - High-level model -> ERD
  - Middle-level model -> DIS (Data Item Set)
  - Low-level model -> Physical model
Three models

- ERD and DIS (detail diagram of ERD)

Subject 1

Subject 2

Subject 3

Subject 4

ERD

DIS

DIS

DIS

Middle-level model

High-level model
Three models

4 groups in a DIS (Customer subject)

Connection data group
- Product ID

Basic data group
- Customer ID
- Name
- Sex
- ID
- Address
- Tel
- E-mail

Class data group
- Transaction ID
- Amount
- Date

2nd level data group
- Transaction ID
- Amount
- Date

Electronic

Furniture

Food
Conceptual model

ERD conceptual model design process

Tasks & environment evaluation

Requirement collection & analysis

Subject selection & their relationship

Subject & content description

ERD

What's in current DB?
How are they organized?
How are they distributed?

Decision types,
Decision interest

Pick the most important subject?
easier to build subject?
Conceptual model

Prescription conceptual model (ERD)

- Disease subject
- Drug subject
- Prescription subject

Disease
  - Disease ID
    - Disease info

Drug
  - Drug basic
    - Drug ID

Prescription
  - Prescription ID
    - Prescription
      - Prescription effects
        - Pres basic
          - Pres related
Logical model

- Analyze and select subject fields
- Determine data granularity
- Determine data partition
- Increase export characters
- Define relational model
- Record system definition, …
Physical model

- Determine data storage structure
  - Distributed,
  - Centralized,
  - RAID (redundant array of inexpensive disk),

- Determine index strategy
  - B-tree indexing
  - Bit-wise indexing
  - Generalized index
  - Connection index
Physical model

- Determine data storage strategy
  - Merge table
  - Data sequence
  - Row-column table
  - Partition table

- Storage optimization
  - Determine the block size
  - Determine block parameters
    - Block percent free, block percent used
Dimensionality modeling

- Multidimensional Data Model
- Dimensionality modeling
  - Star schema, snowflake schema, starflake schema
  - Fact constellation
Multidimensional Data Model

- Multi-dimensional Data as Three-field table versus Two-dimensional Matrix

<table>
<thead>
<tr>
<th>City</th>
<th>Time</th>
<th>Total Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glasgow</td>
<td>Q1</td>
<td>29726</td>
</tr>
<tr>
<td>Glasgow</td>
<td>Q2</td>
<td>30443</td>
</tr>
<tr>
<td>Glasgow</td>
<td>Q3</td>
<td>30582</td>
</tr>
<tr>
<td>Glasgow</td>
<td>Q4</td>
<td>31390</td>
</tr>
<tr>
<td>London</td>
<td>Q1</td>
<td>43555</td>
</tr>
<tr>
<td>London</td>
<td>Q2</td>
<td>48244</td>
</tr>
<tr>
<td>London</td>
<td>Q3</td>
<td>56222</td>
</tr>
<tr>
<td>London</td>
<td>Q4</td>
<td>45632</td>
</tr>
<tr>
<td>Aberdeen</td>
<td>Q1</td>
<td>53210</td>
</tr>
<tr>
<td>Aberdeen</td>
<td>Q2</td>
<td>34567</td>
</tr>
<tr>
<td>Aberdeen</td>
<td>Q3</td>
<td>45677</td>
</tr>
<tr>
<td>Aberdeen</td>
<td>Q4</td>
<td>50056</td>
</tr>
<tr>
<td>..........</td>
<td>....</td>
<td>...............</td>
</tr>
<tr>
<td>..........</td>
<td>....</td>
<td>...............</td>
</tr>
<tr>
<td>..........</td>
<td>....</td>
<td>...............</td>
</tr>
<tr>
<td>..........</td>
<td>....</td>
<td>...............</td>
</tr>
</tbody>
</table>

City

Quarter

Time

<table>
<thead>
<tr>
<th>City</th>
<th>Glasgow</th>
<th>London</th>
<th>Aberdeen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>29726</td>
<td>43555</td>
<td>53210</td>
</tr>
<tr>
<td>Q2</td>
<td>30443</td>
<td>48244</td>
<td>34567</td>
</tr>
<tr>
<td>Q3</td>
<td>30582</td>
<td>56222</td>
<td>45677</td>
</tr>
<tr>
<td>Q4</td>
<td>31390</td>
<td>45632</td>
<td>50056</td>
</tr>
</tbody>
</table>
Multidimensional Data Model

- A table with two dimensions

<table>
<thead>
<tr>
<th>time (quarter)</th>
<th>location = “Vancouver”</th>
<th>item (type)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>home entertainment</td>
<td>computer</td>
</tr>
<tr>
<td>Q1</td>
<td>605</td>
<td>825</td>
</tr>
<tr>
<td>Q2</td>
<td>680</td>
<td>952</td>
</tr>
<tr>
<td>Q3</td>
<td>812</td>
<td>1023</td>
</tr>
<tr>
<td>Q4</td>
<td>927</td>
<td>1038</td>
</tr>
</tbody>
</table>
Multidimensional Data Model

- Multi-dimensional Data as Four-field Table versus Three-dimensional Cube

<table>
<thead>
<tr>
<th>Property Type</th>
<th>City</th>
<th>Time</th>
<th>Total Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>Glasgow</td>
<td>Q1</td>
<td>15056</td>
</tr>
<tr>
<td>House</td>
<td>Glasgow</td>
<td>Q1</td>
<td>14670</td>
</tr>
<tr>
<td>Flat</td>
<td>Glasgow</td>
<td>Q2</td>
<td>14555</td>
</tr>
<tr>
<td>House</td>
<td>Glasgow</td>
<td>Q2</td>
<td>15888</td>
</tr>
<tr>
<td>Flat</td>
<td>Glasgow</td>
<td>Q3</td>
<td>14578</td>
</tr>
<tr>
<td>House</td>
<td>Glasgow</td>
<td>Q3</td>
<td>16004</td>
</tr>
<tr>
<td>Flat</td>
<td>Glasgow</td>
<td>Q4</td>
<td>15890</td>
</tr>
<tr>
<td>House</td>
<td>Glasgow</td>
<td>Q4</td>
<td>15500</td>
</tr>
<tr>
<td>Flat</td>
<td>London</td>
<td>Q1</td>
<td>19678</td>
</tr>
<tr>
<td>House</td>
<td>London</td>
<td>Q1</td>
<td>23877</td>
</tr>
<tr>
<td>Flat</td>
<td>London</td>
<td>Q2</td>
<td>19567</td>
</tr>
<tr>
<td>House</td>
<td>London</td>
<td>Q2</td>
<td>28677</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multidimensional Data Model

- A table with three dimensions

```
<table>
<thead>
<tr>
<th>location = “Chicago”</th>
<th>location = “New York”</th>
<th>location = “Toronto”</th>
<th>location = “Vancouver”</th>
</tr>
</thead>
<tbody>
<tr>
<td>item</td>
<td>item</td>
<td>item</td>
<td>item</td>
</tr>
<tr>
<td>time</td>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>ent.</td>
<td>comp.</td>
<td>comp.</td>
<td>comp.</td>
</tr>
<tr>
<td>phone</td>
<td>sec.</td>
<td>phone</td>
<td>sec.</td>
</tr>
<tr>
<td>Q1</td>
<td>854 882 89 623</td>
<td>1087 968 38 872</td>
<td>818 746 43 591</td>
</tr>
<tr>
<td>Q2</td>
<td>943 890 64 698</td>
<td>1130 1024 41 925</td>
<td>894 769 52 682</td>
</tr>
<tr>
<td>Q3</td>
<td>1032 924 59 789</td>
<td>1034 1048 45 1002</td>
<td>940 795 58 728</td>
</tr>
<tr>
<td>Q4</td>
<td>1129 992 63 870</td>
<td>1142 1091 54 984</td>
<td>978 864 59 784</td>
</tr>
</tbody>
</table>
```
Multidimensional Data Model

- The attributes of a dimension may be related via a hierarchy.

Dimensions: Product, City, Date
Hierarchical summarization paths

- Industry
  - Category
    - Product
- Country
  - State
- Year
  - Quarter
    - Month
    - Week
  - Date
Multidimensional Data Model

A sample data cube
Multidimensional Data Model

- Star schema
  - A schema for multidimensional data model
  - Compare with traditional ER model
Multidimensional Data Model

- Star schema
  - Fact table and dimension tables
Multidimensional Data Model

- Star schema: sales

![Diagram of a star schema with dimensions and facts]
Multidimensional Data Model

- Fact table
  - Quantifies the data described by the Dimension tables
  - Key made up of unique combination of values of dimension keys
    - Always contains date or date dimension
  - Fact values should be additive only
    - Aggregation of quantities or amounts from atomic level
    - No percentages or ratios
Multidimensional Data Model

- Dimension table
  - Describes the data organized in the Fact table
  - Key must be a unique value
  - Key should be the most detailed aggregation level required, if possible (e.g., ZIP code)
  - Manageable number of aggregation levels
  - Dimension does not have to be a hierarchy; can be a combination of attributes
Multidimensional Data Model

Relationship of a star schema and a report
Dimensionality modeling

- A logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access.

- Uses the concepts of Entity-Relationship modeling with some important restrictions.

- Every dimensional model (DM) is composed of one table with a composite primary key, called the fact table, and a set of smaller tables called dimension tables.
Dimensionality modeling

- Each dimension table has a simple (non-composite) primary key that corresponds exactly to one of the components of the composite key in the fact table.

- Forms ‘star-like’ structure, which is called a star schema or star join.
Dimensionality modeling

- All natural keys are replaced with **surrogate keys**. Means that every join between fact and dimension tables is based on surrogate keys, not natural keys.

- Surrogate keys allows the data in the warehouse to have some independence from the data used and produced by the OLTP systems.
Star schema for property sales of **DreamHome**
Dimensionality modeling

- Star schema is a logical structure that has a fact table containing factual data in the center, surrounded by dimension tables containing reference data, which can be de-normalized.

- Facts are generated by events that occurred in the past, and are unlikely to change, regardless of how they are analyzed.
Dimensionality modeling

- Bulk of data in data warehouse is in **fact tables**, which can be extremely large.
- Important to treat fact data as read-only reference data that will not change over time.
- Most useful fact tables contain one or more numerical measures, or ‘facts’ that occur for each record and are numeric and additive.
Dimensionality modeling

- **Dimension tables** usually contain descriptive textual information.
- Dimension attributes are used as the constraints in data warehouse queries.
- Star schemas can be used to speed up query performance by de-normalizing reference information into a single dimension table.
Dimensionality modeling

- **Snowflake** schema is a variant of the star schema where dimension tables do not contain de-normalized data.

- **Starflake** schema is a hybrid structure that contains a mixture of star (denormalized) and snowflake (normalized) schemas. Allows dimensions to be present in both forms to cater for different query requirements.
Property sales with normalized version of Branch dimension table

Dimension Tables
- Branch
  - branchID (PK)
  - branchNo
  - branchType
  - city (FK)
- City
  - city (PK)
  - region (FK)
- Region
  - region (PK)
  - country

Fact Table
PropertySale
- timeID (FK)
- propertyID (FK)
- branchID (FK)
- clientId (FK)
- promotionID (FK)
- staffID (FK)
- ownerID (FK)
- offerPrice
- sellingPrice
- saleCommission
- saleRevenue
Fact and dimension tables for each business process of DreamHome

<table>
<thead>
<tr>
<th>Business process</th>
<th>Fact table</th>
<th>Dimension tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Property sales</td>
<td>PropertySale</td>
<td>Time, Branch, Staff, PropertyForSale, Owner, ClientBuyer, Promotion</td>
</tr>
<tr>
<td>Property rentals</td>
<td>Lease</td>
<td>Time, Branch, Staff, PropertyForRent, Owner, ClientRenter, Promotion</td>
</tr>
<tr>
<td>Property viewing</td>
<td>PropertyViewing</td>
<td>Time, Branch, PropertyForSale, PropertyForRent, ClientBuyer, ClientRenter</td>
</tr>
<tr>
<td>Property advertising</td>
<td>Advert</td>
<td>Time, Branch, PropertyForSale, PropertyForRent, Promotion, Newspaper</td>
</tr>
<tr>
<td>Property maintenance</td>
<td>PropertyMaintenance</td>
<td>Time, Branch, Staff, PropertyForRent</td>
</tr>
</tbody>
</table>