Introduction to Data Warehousing

- Why data warehousing?
- What is data warehousing?
- Characteristics of data warehouse
- Processes & steps in building data warehouse
- Analyzing the content of data warehouse
Why Data Warehousing?

- **Data explosion** in data base management systems (DBMS)
  - Inefficient retrieval of required information
- Needs of Decision Support Systems (DSS) to **facilitate decision making**
  - Extracting, cleaning, transforming, and filtering data from DBMS and provide efficient access to required information
- Data warehouse comes to rescue
Why Data Warehousing?

Who needs data warehouse?

- **Decision makers** who rely on mass amount of data
- Those who use customized, complex processes to obtain information from various data sources
- Those who want to use simple technology to access data
- Those who require systematic approach for decision
Why Data Warehousing?

- Two major functions of data warehousing
  - Extracting necessary information for decision making from heterogeneous data sources and stored in the data warehouse
  - Providing queries and decision analyses to users
Why Data Warehousing?

Wal-Mart deployed data warehousing technology early.

- Started in early 1980s
- 12 GB in 1988
- 24 GB in 1989
- 7.5 TB in 1996
- 24 TB in 1997
- Apply market basket analysis successfully
Why Data Warehousing?

Typical DW Queries

- What was the total revenue for Scotland in the third quarter of 2004?
- What was the total revenue for property sales for each type of property in Great Britain in 2003?
- What are the three most popular areas in each city for the renting of property in 2004 and how does this compare with the figures for the previous two years?
## DreamHome Rental Database

### Figure 3.3
Instance of the DreamHome rental database.

#### Branch

<table>
<thead>
<tr>
<th>branchNo</th>
<th>street</th>
<th>city</th>
<th>postcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>B003</td>
<td>22 Deer Rd</td>
<td>London</td>
<td>SW1 4EH</td>
</tr>
<tr>
<td>B007</td>
<td>16 Argyll St</td>
<td>Aberdeen</td>
<td>AB2 3SU</td>
</tr>
<tr>
<td>B003</td>
<td>163 Main St</td>
<td>Glasgow</td>
<td>G11 9QX</td>
</tr>
<tr>
<td>B004</td>
<td>32 Manse Rd</td>
<td>Bristol</td>
<td>BS9 1NZ</td>
</tr>
<tr>
<td>B002</td>
<td>56 Clover Dr</td>
<td>London</td>
<td>NW10 6EU</td>
</tr>
</tbody>
</table>

#### Staff

<table>
<thead>
<tr>
<th>staffNo</th>
<th>fName</th>
<th>lName</th>
<th>position</th>
<th>sex</th>
<th>DOB</th>
<th>salary</th>
<th>branchNo</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL21</td>
<td>John</td>
<td>White</td>
<td>Manager</td>
<td>M</td>
<td>1-Oct-45</td>
<td>30000</td>
<td>B005</td>
</tr>
<tr>
<td>SG37</td>
<td>Ann</td>
<td>Beec</td>
<td>Assistant</td>
<td>F</td>
<td>10-Nov-60</td>
<td>12000</td>
<td>B003</td>
</tr>
<tr>
<td>SG14</td>
<td>David</td>
<td>Ford</td>
<td>Supervisor</td>
<td>M</td>
<td>24-Mar-58</td>
<td>18000</td>
<td>B003</td>
</tr>
<tr>
<td>SA9</td>
<td>Mary</td>
<td>Howe</td>
<td>Assistant</td>
<td>F</td>
<td>19-Feb-70</td>
<td>6000</td>
<td>B007</td>
</tr>
<tr>
<td>SG5</td>
<td>Susan</td>
<td>Brand</td>
<td>Manager</td>
<td>F</td>
<td>3-Jun-40</td>
<td>24000</td>
<td>B003</td>
</tr>
<tr>
<td>SL41</td>
<td>Julie</td>
<td>Lee</td>
<td>Assistant</td>
<td>F</td>
<td>13-Jun-65</td>
<td>9000</td>
<td>B005</td>
</tr>
</tbody>
</table>

#### PropertyForRent

<table>
<thead>
<tr>
<th>propertyNo</th>
<th>street</th>
<th>city</th>
<th>postcode</th>
<th>type</th>
<th>rooms</th>
<th>rent</th>
<th>ownerNo</th>
<th>staffNo</th>
<th>branchNo</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA14</td>
<td>16 Holhead</td>
<td>Aberdeen</td>
<td>AB7 5SU</td>
<td>House</td>
<td>6</td>
<td>650</td>
<td>CO46</td>
<td>SL9</td>
<td>B007</td>
</tr>
<tr>
<td>PI94</td>
<td>6 Argyll St</td>
<td>London</td>
<td>NW2</td>
<td>Flat</td>
<td>3</td>
<td>400</td>
<td>CO87</td>
<td>SL41</td>
<td>B005</td>
</tr>
<tr>
<td>PG4</td>
<td>6 Lawrence St</td>
<td>Glasgow</td>
<td>G11 9QX</td>
<td>Flat</td>
<td>3</td>
<td>350</td>
<td>CO40</td>
<td>SL9</td>
<td>B003</td>
</tr>
<tr>
<td>PG36</td>
<td>2 Manor Rd</td>
<td>Glasgow</td>
<td>G52 4QX</td>
<td>Flat</td>
<td>6</td>
<td>450</td>
<td>CO93</td>
<td>SG37</td>
<td>B003</td>
</tr>
<tr>
<td>PG21</td>
<td>18 Dale Rd</td>
<td>Glasgow</td>
<td>G12</td>
<td>House</td>
<td>600</td>
<td>CO93</td>
<td>SG37</td>
<td>B003</td>
<td></td>
</tr>
<tr>
<td>PG16</td>
<td>5 Nover Dr</td>
<td>Glasgow</td>
<td>G12 9AX</td>
<td>Flat</td>
<td>4</td>
<td>450</td>
<td>CO93</td>
<td>SG14</td>
<td>B003</td>
</tr>
</tbody>
</table>

#### Client

<table>
<thead>
<tr>
<th>clientNo</th>
<th>fName</th>
<th>lName</th>
<th>telNo</th>
<th>prefType</th>
<th>maxRent</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR76</td>
<td>John</td>
<td>Kay</td>
<td>0207-774-5632</td>
<td>Flat</td>
<td>425</td>
</tr>
<tr>
<td>CR86</td>
<td>Aline</td>
<td>Stewart</td>
<td>0141-848-1825</td>
<td>Flat</td>
<td>350</td>
</tr>
<tr>
<td>CR74</td>
<td>Mike</td>
<td>Ritchie</td>
<td>01475-392178</td>
<td>House</td>
<td>750</td>
</tr>
<tr>
<td>CR62</td>
<td>Mary</td>
<td>Tregear</td>
<td>01224-196720</td>
<td>Flat</td>
<td>600</td>
</tr>
</tbody>
</table>

#### PrivateOwner

<table>
<thead>
<tr>
<th>ownerNo</th>
<th>fName</th>
<th>lName</th>
<th>address</th>
<th>telNo</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO46</td>
<td>Joe</td>
<td>Keogh</td>
<td>2 Fergus Dr, Aberdeen AB2 2XS</td>
<td>01224-862122</td>
</tr>
<tr>
<td>CO87</td>
<td>Carol</td>
<td>Faircl</td>
<td>6 Ashray St, Glasgow G32 9DZ</td>
<td>0141-357-7419</td>
</tr>
<tr>
<td>CO40</td>
<td>Tina</td>
<td>Murphy</td>
<td>63 Well St, Glasgow G42</td>
<td>0141-943-1728</td>
</tr>
<tr>
<td>CO93</td>
<td>Tony</td>
<td>Shaw</td>
<td>12 Park Pl, Glasgow G46QR</td>
<td>0141-225-5025</td>
</tr>
</tbody>
</table>

#### Viewing

<table>
<thead>
<tr>
<th>clientNo</th>
<th>propertyNo</th>
<th>viewDate</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR56</td>
<td>PA14</td>
<td>24-May-04</td>
<td>too small</td>
</tr>
<tr>
<td>CR76</td>
<td>PG4</td>
<td>20-Apr-04</td>
<td>too remote</td>
</tr>
<tr>
<td>CR56</td>
<td>PG4</td>
<td>26-May-04</td>
<td>too small</td>
</tr>
<tr>
<td>CR62</td>
<td>PA14</td>
<td>14-May-04</td>
<td>no dining room</td>
</tr>
<tr>
<td>CR56</td>
<td>PG36</td>
<td>28-Apr-04</td>
<td></td>
</tr>
</tbody>
</table>

#### Registration

<table>
<thead>
<tr>
<th>clientNo</th>
<th>branchNo</th>
<th>staffNo</th>
<th>dateJoined</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR76</td>
<td>B005</td>
<td>SL41</td>
<td>2-Jan-04</td>
</tr>
<tr>
<td>CR56</td>
<td>B003</td>
<td>SG37</td>
<td>11-Apr-03</td>
</tr>
<tr>
<td>CR74</td>
<td>B003</td>
<td>SG37</td>
<td>16-Nov-02</td>
</tr>
<tr>
<td>CR62</td>
<td>B007</td>
<td>SA9</td>
<td>7-Mar-03</td>
</tr>
</tbody>
</table>
Why Data Warehousing?

Typical DW Queries

- What is the **monthly revenue** for property sales at each branch office, compared with rolling 12-monthly prior figures?
- What would be the **effect on property sales** in the different regions of Britain if legal costs went up by 3.5% and Government taxes went down by 1.5% for properties over £100,000?
- Which type of property sells for prices **above the average selling price** for properties in the main cities of Great Britain and how does this correlate to demographic data?
- What is the **relationship** between the total annual revenue generated by each branch office and the total number of **sales staff** assigned to each branch office?
What is Data Warehousing?

A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **non-volatile** collection of data in **support of** management’s **decision-making process** (Inmon, 1993).
What is Data Warehousing?

subject-oriented

- The warehouse is organized around the major subjects of the enterprise (e.g. customers, products, and sales) rather than the major application areas (e.g. customer invoicing, stock control, and product sales).

- This is reflected in the need to store decision-support data rather than application-oriented data.
What is Data Warehousing?

base customer data 1985–1987

customer ID
from date
to date
name
address
phone
dob
sex

............

customer activity
detail 1987–1989

customer ID
activity date
amount
location
for item
invoice no
clerk ID
order no

............

base customer data 1988–1990

customer ID
from date
to date
name
address
credit rating
employer
dob
sex

............

customer activity
detail 1990–1991

customer ID
activity date
amount
location
order no
line item no
sales amount
invoice no
deliver to

.............

customer ID
month
number of transactions
average tx amount
tx high
tx low
txs cancelled

....................
What is Data Warehousing?

Integrated

- The data warehouse integrates corporate application-oriented data from different source systems, which often includes data that is inconsistent.
- The integrated data source must be made consistent to present a unified view of the data to the users.
What is Data Warehousing?
What is Data Warehousing?

- Time-variant data
  - Data in the warehouse is only accurate and valid at some point in time or over some time interval.
  - Time-variance is also shown in the extended time that the data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots.
What is Data Warehousing?

- **operational**
  - time horizon—current to 60–90 days
  - update of records
  - key structure may/may not contain an element of time

- **data warehouse**
  - time horizon—5–10 years
  - sophisticated snapshots of data
  - key structure contains an element of time
What is Data Warehousing?

- non-volatile
  - Data in the warehouse is not updated in real-time but is refreshed from operational systems on a regular basis.
  - New data is always added as a supplement to the database, rather than a replacement.
What is Data Warehousing?

- **Operational Data**
  - Insert (isrt)
  - Delete (dlet)
  - Change (chng)
  - Access (access)

- **Data Warehouse**
  - Load (load)
  - Access (access)

Record-by-record manipulation of data

Mass load/access of data
What is Data Warehousing?

Data in the data warehouse

- Query responses
- Aggregated data
- Primitive data
- Processing data

Meta data

- Data describing data in the warehouse
  - Structure, content, keys, indices
  - Like data dictionary in the DBMS
Characteristics of data warehouse

- Subject & subject-oriented
- Integration
- Non-updatable
- Invariant
Processes & steps in building data warehouse

- Processes
  - Extraction
  - Transformation
  - Cleaning
  - Loading
  - Aggregate
Processes & steps in building data warehouse

Basic steps in building data warehouse

1. Collect & analyze business requirement
2. Build data model and data warehouse physical design
3. Define data sources
4. Select data warehouse technology & platform
5. Extract, transform, clean & load data into the warehouse
Processes & steps in building data warehouse (con’t)

Basic steps in building data warehouse

6. Select access and report tools
7. Select database access tools
8. Select data analysis & display tools
9. Update data warehouse
Analyzing the content of data warehouse

- On-Line Analytical Processing (OLAP)
  - Multidimensional data analysis (MDA)
  - Data Cube (DC)
- Data Mining (DM)
## Comparison of OLTP Systems and Data Warehousing

<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>Ad hoc, 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>
Homework

- What are the difference between transaction data and analysis data?
- What is data warehouse and characteristic?
- What is the procedure of entering data into a data warehouse?
- Compare the data warehouse user and non-data warehouse user.