What is Data Warehouse?

• Defined in many different ways, but not rigorously.
  – A decision support database that is maintained separately from the organization’s operational database
  – Support information processing by providing a solid platform of consolidated, historical data for analysis.
  – Provides tools for business executives to systematically organize and use the data to make strategic decisions

• “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
Data Warehouse—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.
Data Warehouse—Integrated

• Constructed by integrating multiple, heterogeneous data sources
  – relational databases, flat files, on-line transaction records
• Data cleaning and data integration techniques are applied.
  – 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.
Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)

- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”.

Data Warehouse—Non-Volatile

• 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.
From Tables and Spreadsheets to Data Cubes

• A data warehouse is based on a multidimensional data model which views data in the form of a data cube.
• A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
• It is defined by dimensions and facts
  – Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  – Fact table contains measures (such as dollars_sold) and keys to each of the related dimension table
Multidimensional Data

- Sales volume as a function of product, month, and region

Dimensions: Product, Location, Time
Hierarchical summarization paths

Industry  Region  Year
Category   Country Quarter
Product   City Month Week
          Office Day
4-D Cube (Supplier dimension)

Figure 2.2 A 4-D data cube representation of sales data, according to the dimensions time, item, location, and supplier. The measure displayed is dollars_sold (in thousands). For improved readability, only some of the cube values are shown.
## A Sample Data Cube

<table>
<thead>
<tr>
<th>Product</th>
<th>TV</th>
<th>PC</th>
<th>VCR</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>1Qtr</td>
<td>2Qtr</td>
<td>3Qtr</td>
<td>4Qtr</td>
</tr>
<tr>
<td>U.S.A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canada</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Total annual sales of TV in U.S.A.
Cuboids

• In data warehousing literature, an n-D base cube is called a **base cuboid**. The top most 0-D cuboid, which holds the highest-level of summarization, is called the **apex cuboid**. The lattice of cuboids forms a **data cube**.
Cuboids Corresponding to the Cube

- **0-D (apex) cuboid**
- **1-D cuboids**
- **2-D cuboids**
- **3-D (base) cuboid**
Browsing a Data Cube

- Visualization
- OLAP capabilities
- Interactive manipulation
Cube: A Lattice of Cuboids

0-D (apex) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (base) cuboid