Data Warehousing

Organization of Concepts

Large-Scale Data Management for Organizational Decision-Making

- **Data Warehouse (Data Integration)**
  - Purpose
  - Basic Features
    - Definition
    - Data Requirements
    - Transaction Processing vs. Data Warehousing
    - Data Structures
    - Example
  - Architecture
    - Data Mapping
    - Example Architecture of an IBM Data Warehouse
  - Preparing Data for DW
  - Reasons for Failure

- **Multidimension Databases (Enabling Technology)**
  - Purpose
  - Basic Features
    - MDDB vs. Relational Database
    - Rotation
    - Ranging
    - Roll-Up
    - Drill-Down
    - Computations
  - Benefits
  - Tools
    - PowerPlay

- **Data Mining (Relationship Discovery)**
  - Purpose
  - Applications
  - Functions
    - Association
    - Availability Bias
    - Sequential Pattern
    - Classifying
    - Clustering
    - Hybrid
  - Emergent Applications

Large-Scale Data Management for Organizational Decision-Making
Learning Objectives

59. Explain what is a Data-Warehouse (DW) and the rationale for building them.
60. Describe four levels of data granularity handled by a DW.
61. Identify data requirements for DW and Database Technology suitable for such requirements
62. Explain what is metadata and its importance for DW.
63. Describe the process of getting data into DW.
64. Identify common reasons for DW failure.

Appendix 1
- Architecture of a Data Warehouse: An Example from IBM
- Data Mapping: An Example from IBM

Definition: Data-Warehouse (L.O.59)

- A data warehouse is a *central source* of data, stocked with data extracted from different operational systems and standardized.
  - New hybrid information systems which bridge two broad classes of systems, namely: storage and retrieval systems, and decision support systems
  - The objective is not just the storage and retrieval of data objects, but also the programmed and ad-hoc transformation of data to information for its defined user base.
Illustration (L.O.59)

1. Some organizations, such as Grand Metropolitan PLC, consider their data warehouse essentially as front-ends to existing production systems.

2. Other organizations such as Owens-Corning Fiberglass Corp. view their data warehouse as a collection of subject- or application-oriented databases that are kept separate from the operational applications.

Rationale for DW (Why?)
Reason #1: Data Resource Bottleneck (L.O.59)

• Knowledge workers need to gather and analyze business data to generate business information
• “Data in jail” syndrome
• Data quality problems, including:
  – missing data
  – unreliable data
  – invalid data
  – inappropriate granularity
  – problematic semantics
Rationale for DW (Why?)
Reason #2: Conceptual Fit for End Users Use (L.O.59)

• Most organizational transaction systems designed for capture and storage, but not amenable for analysis
• Offers a good conceptual fit with the way end-users visualize business data
  – Most business people already think about their businesses in multidimensional terms
    • Managers tend to ask questions about product sales in different markets over specific time periods

Reason #3: The New Technological Environment (L.O.59)

Enterprise Network Computing and Client/Server Technology are changing the way organizations look at all of their information systems
Reason #3 continues: Implications of Newer Technological Capabilities (L.O.59)

- Emergence of Tools
  - Emergence of large-scale data management tools
  - Emergence of several end-user oriented data exploration tools
- Support for increased flexibility and integration within environments
  - Transaction environments
  - Decision support environment
- Support for increased flexibility and integration between environments
  - Improved interface between transactions and decision support.

Reason #4: Overall Goals of DataWarehousing (L.O.59)

- **Performance** (Canned queries, MD Analysis, Ad hoc, Min. Impact on Operational System)
- **Flexibility** (MD Flex, Ad hoc, Change data structure)
- **Scalability** (No. of Users, Volume of Data)
- **Ease of Use** (Location, Formulation, Navigation, Manipulation)
- **Data Quality** (Consistent, Correct, Timely, Integrated)
What types of data DW deals with?

Four degrees of Data Granularity (L.O.60)

- **Feed level:**
  1. **Operational** data from TPS - fundamental to DW data
     - Ex. sales of XYZ at the end of today is 2345 units

- **Storage level:**
  2. **Atomic** data - DW data - individual data item and mostly the lowest level of data in DW. Ex. sales of XYZ at the end of Sept. is 2045 units
     - … at the end of August 1743
     - … at the end of July 1345, etc…
  3. **Summary** data - DW data - these summaries are calculated ahead of time and stored for recall. Ex. At the end of third quarter sales of XYZ for Southeastern region was 23,456; at the end of second quarter sales of XYZ…
  4. **Answers to specific question** - processed and stored for individual user’s pc or workstation.
     - How does the growth of XYZ sales in SE compares with NW?

DW data requirements (L.O.61)

- DW data requirements are different from traditional transaction processing (TP) data. For example:

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>TP data</th>
<th>DW data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volatility</td>
<td>Dynamic</td>
<td>Static</td>
</tr>
<tr>
<td>Currency</td>
<td>Current</td>
<td>Historical</td>
</tr>
<tr>
<td>Time dimension</td>
<td>Implicitly “now”</td>
<td>Explicit, visible</td>
</tr>
<tr>
<td>Granularity</td>
<td>Primitive, detailed</td>
<td>Detailed and derived summaries</td>
</tr>
<tr>
<td>Updates</td>
<td>Continuous, random</td>
<td>Periodic, scheduled</td>
</tr>
<tr>
<td>Tasks</td>
<td>Repetitive</td>
<td>Unpredictable</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Performance</td>
<td>High performance mandatory</td>
<td>Lower performance may be acceptable</td>
</tr>
</tbody>
</table>
Database Technology suitable for the above DW data requirements (L.O.61)

• Hierarchical and Network structures based technologies are not suitable
• Simple Relational with some ‘fixing’ can do the job
• A new structure - which is quite unsuitable for TP data - has emerged for DW use: “Multidimensional DB”
  – Time, itself, is a dimension in MDDB

Database Technologies for DW continues...

• All DWs have two types of data
  – Dimension data (also called Perspectives)
    • Tens to a few mil. Rows
    • Textually described
    • Frequently modified
  – Fact data (also called Measurements)
    • Millions or billions of rows
    • Numeric
    • Don’t change
Pet store example

- Multi-dimensional Database Structure

<table>
<thead>
<tr>
<th></th>
<th>Cats</th>
<th>Dogs</th>
<th>Fish</th>
<th>Gerbils</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>3</td>
<td>12</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Baker</td>
<td>2</td>
<td>4</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>Carstens</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>16</td>
</tr>
</tbody>
</table>

Dimension data: "Salesperson"
Dimension data: "Pets"

Fact data

Relational Equivalence (L.O 61 finishes here)

<table>
<thead>
<tr>
<th>Salesperson</th>
<th>Pet Type</th>
<th>No. Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>Cats</td>
<td>3</td>
</tr>
<tr>
<td>Adams</td>
<td>Dogs</td>
<td>12</td>
</tr>
<tr>
<td>Adams</td>
<td>Fish</td>
<td>5</td>
</tr>
<tr>
<td>Adams</td>
<td>Gerbils</td>
<td>4</td>
</tr>
<tr>
<td>Baker</td>
<td>Cats</td>
<td>2</td>
</tr>
<tr>
<td>Baker</td>
<td>Dogs</td>
<td>4</td>
</tr>
<tr>
<td>Baker</td>
<td>Fish</td>
<td>22</td>
</tr>
<tr>
<td>Baker</td>
<td>Gerbils</td>
<td>5</td>
</tr>
<tr>
<td>Carstens</td>
<td>Cats</td>
<td>6</td>
</tr>
<tr>
<td>Carstens</td>
<td>Dogs</td>
<td>2</td>
</tr>
<tr>
<td>Carstens</td>
<td>Fish</td>
<td>3</td>
</tr>
<tr>
<td>Carstens</td>
<td>Gerbils</td>
<td>16</td>
</tr>
</tbody>
</table>
More on MDDB later in the lecture

Implications due to Metadata and Semantics
(L.O.62)

– Human
  • “Natural” assumptions are not always correct
  • What to do with same item - shoe size - length vs. width
  • Is ‘pay-rate’ same as ‘salary’ in another database
  • Validation rules for common human items such as U.S. zip code

– Computer-based metadata for people to use
  • create metadata warehouse
    – data descriptions, data structure
    – A business thesaurus; etc.

– Computer-based metadata for computer to use
Getting Data into the Data Warehouse (L.O.62)  
(Self study items)

• Extraction -- extract data
• Cleansing -- to achieve consistent format, valid values
• Loading -- input data from multiple sources; decide which set of data “rules”
• Transformation
  – Change data type of field from integer to character
  – Summarization -- aggregation

Example of Summarization (L.O.62)

<table>
<thead>
<tr>
<th></th>
<th>Cats</th>
<th>Dogs</th>
<th>Fish</th>
<th>Gerbils</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>3</td>
<td>12</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Baker</td>
<td>2</td>
<td>4</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>Carstens</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>18</td>
<td>30</td>
<td>25</td>
</tr>
</tbody>
</table>

Four cells added for summaries by pet.

Summary data
Why do Data Warehousing Projects Fail? (L.O.64)

(Reference: Anatomy of a Failure
http://www.cio.com/archive/enterprise/111597_data.html)

Summary/ Review Questions for Data Warehouse

• What is a data warehouse (DW)? Why build it? What are its goals?
• DW handles data of different degrees of ‘granularity.’ Do you know what are the levels of data handled by DW?
• What is the relationship between transaction processing systems (and, its data) and a DW? How do their data requirements compare?
• Quite a few DW projects fail or finish without achieving its complete set of original objectives. What are the factors that explain such failures? What can be done to improve the chances of success of DW project?
Multi-Dimensional Databases

Learning Objectives

• L.O.65 Explain what are Multidimensional Databases and why are they needed?
• L.O.66 Contrast MDD and Relational Databases
• L.O.67 Provide examples of situations when is MDD (In)appropriate?
• L.O.68 Identify benefits of MDD
• Appendix 2
  – Examples of MDD Features such as Rotation, Ranging, Roll-Up and Drill-Down, Computations.
  – Example of inner workings of a commercially available OLAP (On-Line Analytical Processing) tool.
What is a Multi-Dimensional Database? (L.O.65)

A multidimensional database (MDD) is a computer software system designed to allow for the efficient and convenient storage and retrieval of large volumes of data that is (1) intimately related and (2) stored, viewed and analyzed from different perspectives. These perspectives are called dimensions.
Why Multi-Dimensional Databases? (L.O.65)

- No single "best" data structure for all applications within an enterprise
- Inherent ability to integrate and analyze large volumes of enterprise data
- Offers a good conceptual fit with the way end-users visualize business data
  - Most business people already think about their businesses in multidimensional terms
  - Managers tend to ask questions about product sales in different markets over specific time periods

Contrasting Relational and Multi-Dimensional Models: Example 1 (L.O. 66)

<table>
<thead>
<tr>
<th>MODEL</th>
<th>COLOR</th>
<th>SALES VOLUME</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINI VAN</td>
<td>BLUE</td>
<td>6</td>
</tr>
<tr>
<td>MINI VAN</td>
<td>RED</td>
<td>5</td>
</tr>
<tr>
<td>MINI VAN</td>
<td>WHITE</td>
<td>4</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>BLUE</td>
<td>3</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>RED</td>
<td>5</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>WHITE</td>
<td>5</td>
</tr>
<tr>
<td>SEDAN</td>
<td>BLUE</td>
<td>4</td>
</tr>
<tr>
<td>SEDAN</td>
<td>RED</td>
<td>3</td>
</tr>
<tr>
<td>SEDAN</td>
<td>WHITE</td>
<td>2</td>
</tr>
</tbody>
</table>
Contrasting Relational Model and MDD-
Example 2 (L.O.66)

SALES VOLUMES FOR ALL DEALERSHIPS

<table>
<thead>
<tr>
<th>MODEL</th>
<th>COLOR</th>
<th>DEALERSHIP</th>
<th>VOLUME</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINI VAN</td>
<td>BLUE</td>
<td>CLYDE</td>
<td>6</td>
</tr>
<tr>
<td>MINI VAN</td>
<td>BLUE</td>
<td>GLEASON</td>
<td>6</td>
</tr>
<tr>
<td>MINI VAN</td>
<td>RED</td>
<td>CLYDE</td>
<td>3</td>
</tr>
<tr>
<td>MINI VAN</td>
<td>RED</td>
<td>GLEASON</td>
<td>5</td>
</tr>
<tr>
<td>MINI VAN</td>
<td>WHITE</td>
<td>CLYDE</td>
<td>2</td>
</tr>
<tr>
<td>MINI VAN</td>
<td>WHITE</td>
<td>GLEASON</td>
<td>4</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>BLUE</td>
<td>CLYDE</td>
<td>2</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>BLUE</td>
<td>GLEASON</td>
<td>3</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>RED</td>
<td>CLYDE</td>
<td>7</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>RED</td>
<td>GLEASON</td>
<td>5</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>WHITE</td>
<td>CLYDE</td>
<td>4</td>
</tr>
<tr>
<td>SPORTS COUPE</td>
<td>WHITE</td>
<td>GLEASON</td>
<td>3</td>
</tr>
<tr>
<td>SEDAN</td>
<td>BLUE</td>
<td>CLYDE</td>
<td>6</td>
</tr>
<tr>
<td>SEDAN</td>
<td>BLUE</td>
<td>GLEASON</td>
<td>4</td>
</tr>
<tr>
<td>SEDAN</td>
<td>BLUE</td>
<td>CARR</td>
<td>2</td>
</tr>
<tr>
<td>SEDAN</td>
<td>RED</td>
<td>CLYDE</td>
<td>1</td>
</tr>
<tr>
<td>SEDAN</td>
<td>RED</td>
<td>GLEASON</td>
<td>3</td>
</tr>
<tr>
<td>SEDAN</td>
<td>RED</td>
<td>CARR</td>
<td>4</td>
</tr>
<tr>
<td>SEDAN</td>
<td>WHITE</td>
<td>CLYDE</td>
<td>2</td>
</tr>
<tr>
<td>SEDAN</td>
<td>WHITE</td>
<td>GLEASON</td>
<td>2</td>
</tr>
<tr>
<td>SEDAN</td>
<td>WHITE</td>
<td>CARR</td>
<td>3</td>
</tr>
</tbody>
</table>
Multidimensional Representation

Sales Volumes

- Assume that each dimension has 10 positions, as shown in the cube above.
- How many records would be there in a relational table?
- Implications for viewing data from an end-user standpoint?
Adding Dimensions- An Example

Sales Volumes

Differences between MDD and Relational Databases (L.O.66)

- Multidimensional array structure represents a higher level of organization than the relational table.
- Perspectives are embedded directly into the structure in the multidimensional model.
  - All possible combinations of perspectives containing a specific attribute (the color BLUE, for example) line up along the dimension position for that attribute.
- Perspective values are placed in fields in the relational model - tells us very little about the overall field contents.
Benefits of MDD (L.O. 68)

- Cognitive Advantages for the User
- Ease of Data Presentation and Navigation
  - Obtaining the same views in a relational world requires the end user to either write complex SQL queries or use an SQL generator against the relational database to convert the table outputs into a more intuitive format.
- Ease of Maintenance
  - Because data is stored in the same way as it is viewed (i.e. according to its fundamental attributes), no additional overhead is required to translate user queries into requests for data.
- Performance
  - Multidimensional databases achieve performance levels that are difficult to match in a relational environment.

When is MDD (In)appropriate? (L.O. 67)

- When interrelationships more meaningful than individual data elements themselves.
- The greater the number of inherent interrelationships between the elements of a dataset, the more likely it is that a study of those interrelationships will yield business information of value to the company.
- Highly interrelated dataset types be placed in a multidimensional data structure for greatest ease of access and analysis.
When is MDD Inappropriate? (L.O.67)

First, consider situation 1

<table>
<thead>
<tr>
<th>LASTNAME</th>
<th>EMPLOYEE#</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMITH</td>
<td>01</td>
<td>21</td>
</tr>
<tr>
<td>REGAN</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>Fink</td>
<td>31</td>
<td>68</td>
</tr>
<tr>
<td>WILD</td>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>KELLY</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td>LINK</td>
<td>03</td>
<td>56</td>
</tr>
<tr>
<td>KRANZ</td>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td>LUCUS</td>
<td>33</td>
<td>41</td>
</tr>
<tr>
<td>WEISS</td>
<td>23</td>
<td>19</td>
</tr>
</tbody>
</table>

Then try to set up an MDD structure for situation 1, with LAST NAME and Employee# as dimensions, and AGE as the measurement.

When is MDD (In)appropriate? (L.O.67)

MDD Structures for the Situations

Note the sparseness between the two MDD representations
Summary and Review Questions

• Why is Multi-dimension Data Base used?
• How does a Multi-Dimension Data Base (MDDB) differ from other database techniques such as relational database?
• When is an MDDB appropriate to be used?

• (DW+MDDB+Data Mining) What roles a DW, an MDDB, and Data-mining play for large scale data management? Can you comment on their interdependence?

Appendix 1

– Architecture of a Data Warehouse: An Example from IBM
– Data Mapping: An Example from IBM
Data Mapping Strategy at IBM

- Business segment is defined as any functional area of IBM such as marketing, finance, manufacturing, etc.
- Business processes is any process used by any business segment.
- A business segment owner may own multiple business processes such as software and hardware development.

Appendix 2

- Examples of MDD features
- An OLAP tool
• Also referred to as “data slicing.”
• Each rotation yields a different slice or two dimensional table of data.
MDD Features - Ranging

- The end user selects the desired *positions* along each *dimension*.
- Also referred to as "data dicing."
- The data is scoped down to a subset grouping.

MDD Features - Roll-Ups & Drill Downs

- The figure presents a definition of a *hierarchy* within the organization dimension.
- Aggregations perceived as being part of the same dimension.
- Moving up and moving down levels in a hierarchy is referred to as “roll-up” and “drill-down.”
Rolling-up and Drilling-down Through Multiple Dimensions. In the previous graphic, we saw that users can roll-up and drill-down through a single dimension, ORGANIZATION. Well designed multidimensional databases also allow users to roll-up and drill down through multiple dimensions concurrently. Thus, in this example, an end-user could hold the positions MANAGER, DISTRICT and PRODUCT constant, while drilling-down or rolling-up through sales figures over the TIME dimension.

MDD Features:
*Drill-Down* Through a Dimension

Sales Volumes

- **COLOR**
- **MODEL**
- **REGION**
- **DISTRICT**
- **DEALERSHIP**
MDD Features: Multidimensional Computations

- Well equipped to handle demanding mathematical functions.
- Can treat arrays like cells in spreadsheets. For example, in a budget analysis situation, one can divide the ACTUAL array by the BUDGET array to compute the VARIANCE array.
- Applications based on multidimensional database technology typically have one dimension defined as a "business measurements" dimension.
- Integrates computational tools very tightly with the database structure.

The Time Dimension

- TIME as a predefined hierarchy for rolling-up and drilling-down across days, weeks, months, years and special periods, such as fiscal years.
  - Eliminates the effort required to build sophisticated hierarchies every time a database is set up.
  - Extra performance advantages