Managing Information Resources

MSC 3824
Class Discussion Notes
Session 4: Database Mgt
Professor Yang W. Lee
Learning Objectives

• **Technology**
  database management systems (relational databases)
  ER model and its use for Schema Integration

• **Applications**
  target marketing
  data mining
  information systems analysis
  data quality analysis

• **Management Issues**
  role in e-commerce & traditional business practices
  strategic applications
  implications of outsourcing data management functions
Data Management

COLLECT
- Data Sources
- Represent

STORE
- Databases
- Data Warehouses

MANAGE
- Inhouse vs Outsourcing
- Processing (Transaction versus Web)

USE
- Data Mining
- Target Marketing

Source: 2002 Course Technology, V. Storey,
Data and their Sources

• Types of data
  – public
  – private

• Sources
  – internal
  – external

• Implications?
Databases for e-Commerce

“Without databases e-commerce is nothing.”

--Former dot com executive
e-Commerce architecture:

Role of DB in 3-tier architecture

SERVER ENVIRONMENT

E-Commerce environment

E-Commerce application (e-shop etc)

Database

Web Server

1st Layer: Presentation Layer

Request Info

Receive Info

Client

2nd Layer: Application Layer

3rd Layer: Data Layer

4/15/2003 © 2003 Y. Lee
Databases

• Database --- integrated collection of data

• Database management software --- mechanism for storing and organizing data for sophisticated queries and manipulation of data

• Relational database – most popular, data organized into tables (Microsoft SQL Server, Oracle, DB2)
Database System for Bank

Customer Database (Data)

CUSTOMER DATA
- Cust name
- SSN
- Address
- Savings Account #
- Loan Account #
- Investment Account #

SAVINGS DATA
- Savings Account #
- Account balance

LOAN DATA
- Loan Account #
- Account balance

INVESTMENT DATA
- Investment Account #
- Account balance

Business Application Program (Business Logic)

Data Consumer
(Bank managers and officers)

SAVINGS SYSTEM

LOAN SYSTEM

INVESTMENT SYSTEM

Source: Dorit Nevo
Data Management

- COLLECT
  - Data Sources
  - Represent
    - Entity Relationship Model

- STORE
  - Databases
  - Data Warehouses

- MANAGE
  - Inhouse vs Outsourcing
  - Transaction Processing

- USE
  - Data Mining
  - Target Marketing

Data Sources, Represent, Databases, Data Warehouses, Inhouse vs Outsourcing, Transaction Processing, Data Mining, Target Marketing
ER Model
by Peter Chen

• Structured methodology used in industry
• Case tools available for developers and systems analysts
• De-facto standard for systems analysis and data needs representation
Database Design
Representation
Entity-Relationship Model

**ENTITY**: Person, place, thing, event about which data must be kept

- **ATTRIBUTE**: Description of a particular ENTITY
- **PRIMARY KEY FIELD**: Primary Field used to retrieve, update, sort RECORD

Sources: @2002 Prentice Hall
@ 2002 Course Technology
Primary KEY FIELD

- Field in each record
- **Uniquely Identifies** THIS Record
- For RETRIEVAL
  
  UPDATING
  
  SORTING

Sources: @2002 Prentice Hall
@ 2002 Course Technology
Simple Example:
“Students take courses”
Use Entity-Relationship Model
Typical Queries using SQL

- SELECT attribute, column
- FROM entity, table
- WHERE condition of column

Note: Minimal redundancy in database design
Normalization: 3rd normal form is industry standard
Relational Schema

**DEPT (DEPT#, BUDGET)**
**EMP (SS#, NAME, AGE, DEPT#)**
**PROJECT (PROJ#, PNAME)**
**WORK_FOR (SS#, PROJ#, %TIME)**

Which departments have employees working on project p3?

SQL (Structured Query Language)

```
SELECT EMP.DEPT#
FROM EMP, WORK_FOR
WHERE (WORK_FOR.PROJ# = P3 AND
(WORK_FOR.SS# = EMP.SS#))
```
Database management:
Example functions and concepts

Data Integrity rules
Concurrency control
Locking against “deadly embraced”
Fragmentation: Horizontal and Vertical
Data Management

COLLECT
- Data Sources
- Represent

STORE
- Databases
- Data Warehouses

MANAGE
- Inhouse vs Outsourcing
- Processing (Transaction versus Web)

USE
- Data Mining
- Target Marketing

Schema Integration
Application of ER Model: SCHEMA INTEGRATION

PURPOSES FOR SCHEMA / VIEW INTEGRATION

- **NEW DATABASE DESIGN**: MERGE MULTIPLE PEOPLE’S VIEWS
- **DATA WAREHOUSE**: CREATE INTEGRATED SCHEMA
- **DATABASE REDESIGN**: MERGE MULTIPLE EXISTING DB’S
- **INTEGRATED SCHEMA FOR DISTRIBUTED HETEROGENEOUS DBMS**

Source: Ozsu and Madnick
PROCESS OF SCHEMA INTEGRATION

- DATA FROM AUTONOMOUS SOURCE
- USEFUL DATA CAPTURES “MEANING” OF DATA
- CAPACITY FOR “ENTERPRISE-WIDE” VIEW OF DATA

TERMINOLOGY

- SCHEMA INTEGRATION
  - COMPONENT SCHEMA
  - INTEGRATED SCHEMA
- VIEW INTEGRATION
  - USER VIEW
  - CONCEPTUAL VIEW
- DATABASE INTEGRATION
  - LOCAL SCHEMA
  - GLOBAL SCHEMA

EXAMPLE

- TWO SCHEMAS DEVELOPED BY TWO GROUPS
- BOTH FOCUS ON BOOKS/PUBLICATIONS
ENTITY-RELATION (ER) DIAGRAM

- ENTITIES
- RELATIONSHIPS
- ATTRIBUTES

Source: Ozsu and Madnick
Step 1. Original schemas

Step 2. Choose “Topics” for “keyword” (Schema 2).

Step 3. Make Publisher into an entity (Schema 2).
Step 4. Superimposition of schema

Step 5. Creation of subset relationship
Inheritance

Step 6. Drop the properties of Book common to Publication
Example Conflicts/Differences

• NAMING CONFLICTS
  – HOMONYMS: SAME NAME, DIFFERENT CONCEPT
    E.G.,
    E.G.,
  – SYNONYMS: SAME CONCEPT, DIFFERENT NAMES
    E.G., CLIENT, CUSTOMER, PATIENT, ...

• STRUCTURAL CONFLICTS
  – TYPE: ENTITY VS. ATTRIBUTE (E.G., PUBLISHER)
  – DEPENDENCY: 1:1 VS. N:M
  – KEY: ALTERNATE WAYS TO IDENTIFY ENTITY (E.G., SS# VS. EMP#)
  – BEHAVIORAL: DIFFERENT INSERTION/DELETION POLICIES (E.G., CAN YOU HAVE DEPARTMENT WITH NO EMPLOYEES?)
CAUSES OF DIFFERENCES

• DIFFERENT PERSPECTIVES
  – DIFFERENT NAME FOR SAME CONCEPT
  – INTERVENING STEPS
    • EMPLOYEE-DEPARTMENT
    • EMPLOYEE-PROJECT-DEPARTMENT

• EQUIVALENCE AMONG CONSTRUCTS
  – MODEL AS ATTRIBUTE OR ENTITY (E.G., PUBLISHER)

• INCOMPATIBLE DESIGN SPECIFICATIONS
  – RELATIONSHIP DIFFERENCES
    (E.G., 1:N VS. N:M)
EXAMPLE SCHEMAS TO BE INTEGRATED:

Departmental database

name

DEPT

n

n

STUDENT

PROFESSOR

m

name

position

SUBJECT

Registrar's database

COURSE

sub #

STUDENT

name

grade

name

id

STUDENT

name

id

Staff Telephone directory

Staff

name

room #

ext #

street

city

phone #

Student Telephone directory

Student

name

year

O/I

OFFICE

ADDRESS

room #

ext #

street

city

phone #

HOME

ADDRESS

street

city

phone #

TERM

ADDRESS

street

city

phone #

Source: Madnick
Schema Integration

SUMMARY

• Increasing need to provide an Integrated GLOBAL VIEW of an organization’s information
  - AND SOMETIMES RELATED ORGANIZATIONS (CUSTOMERS / SUPPLIER)

• AN IMPORTANT STEP IS THE CREATION OF A GLOBAL SCHEMA
  - INTEGRATES SEPARATE SCHEMAS
  - CONTAINS ALL THE CRITICAL INFORMATION NEEDED

Source: Ozsu and Madnick
Data Management

COLLECT
- Data Sources
- Represent

STORE
- Databases
- Data Warehouses

MANAGE
- Inhouse vs Outsourcing
- Processing (Transaction versus Web)

USE
- Data Mining
- Target Marketing
Data Warehouses
Now that we have gathered so much data, What do we do with it?

“I never waste memory on things that can easily be stored and retrieved from elsewhere.”

- Albert Einstein
What is a Data Warehouse?

An integrated and consistent store of subject-oriented data, structured for query and retrieval in order to support management decision making.
A **data warehouse** is where the information systems department puts data to be retrieved for further use.

One cannot just dump masses of data into a disk drive and expect it to be usable.
Bigger and Bigger

Megabyte 1 Million Bytes
Gigabtype 1 Billion Bytes
Terabyte 1000 Gigabytes
Petabyte 1000 Terabytes
Exabyte 1000 Petabytes
ZettaByte 1000 Exabytes
Yottabyte 1000 Zettabytes
Data Warehouse Architecture

Data Sources

Operational DBs

Other sources

metadata

Extract Transform Load Refresh

Data Warehouse

Serve

Analysis Query Reports Data mining

Tools

Source: 2002 Course Technology and Storey

© 2003 Y. Lee
Data Warehouses

• A logical collection of information
  – gathered from many different operational databases
  – supports business analysis activities and decision-making tasks.

• Why?
  – Raw data necessary to make sound business decisions stored in variety of locations and formats

Source: Dorit Nevo, UBC
Wal-Mart

• Challenge:
  – Approximately 100,000 items; over 2,900 store locations
  – Tremendous growth resulted in divergent customer base
  – Need to merchandise locally
    • decisions at the individual store level

• Solution:
  – Data warehouse to store detailed sales data (e.g., by item, store, day)
  – Over 24 terabytes of data
  – Company’s 4,000 suppliers have access to warehouse

• Results:
  – Buyers and vendors:
    • analyze sales trends and make informed decisions
    • analyze customer and seasonal buying trends
    • make mark-down decisions
    • react to merchandise volume and movement at any time
  – Improved in-stock percentage by 10 points (to 96%)
    • one of the highest rates in the industry

Source: Dorit Nevo and V. Storey 4/15/2003
Data Warehouses: Concepts

- Information gathered from many databases
- Multidimensional
- Support OLAP, not OLTP
  - support decision making, not transaction processing

Source: Dorit Nevo and V. Storey
Multidimensional View of Data

• Example: Sales volume as a function of product, time, and location

3 Dimensions
Product, City, Year

Source: Dorit Nevo and V. Storey
OLTP versus OLAP

• **Online transaction processing (OLTP)**
  – gathering input information, processing it, and updating existing information to reflect the gathered and processed information.
  – *Operational databases* are databases that support OLTP.

• **Online analytical processing (OLAP)**
  – manipulation of information to support decision making
  – *Data Warehouses* are special form of database that support OLAP

Source: Dorit Nevo, UBC
4/15/2003

© 2003 Y. Lee
## OLTP vs OLAP

<table>
<thead>
<tr>
<th><strong>User</strong></th>
<th><strong>OLTP</strong></th>
<th><strong>OLAP</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Clerk, IT Professional</td>
<td>Knowledge worker</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Function</strong></th>
<th><strong>OLTP</strong></th>
<th><strong>OLAP</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Day to day operations</td>
<td>Decision support</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Data</strong></th>
<th><strong>OLTP</strong></th>
<th><strong>OLAP</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>Historical</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>View</strong></th>
<th><strong>OLTP</strong></th>
<th><strong>OLAP</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Detailed, Relational</td>
<td>Summarized, Multidimensional</td>
<td></td>
</tr>
</tbody>
</table>

Source: Dorit Nevo and V. Storey
Data Warehouse Considerations

- Who will the consumers/users be?
- What will be their mission?
- How up-to-date must the information be?
- What kind of data will they need for their work?
- Are appropriate data mining tools available?
Data Mining

• An information analysis software that automatically analyze large volumes of data to identify patterns, trends, and relationships in a data warehouse.

• Built-in analysis tools such as advanced statistical techniques are included in the data mining software.

• Query tools: supported by data warehouses
  Information discovery tools: supported by data mining

Source: Dorit Nevo, UBC
4/15/2003
<table>
<thead>
<tr>
<th>Organization</th>
<th>Datamining Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Imperial Bank of Commerce (CIBC)</td>
<td>Customer profitability system helps the bank identify its most profitable customers so that it can offer them special sales and services.</td>
</tr>
<tr>
<td>Stein Roe Investors</td>
<td>Analyzes data generated by visitors to its Web site to create profiles of existing and prospective customers. The company can use these profiles to target potential customers with content, advertising and incentives geared to their interests, such as retirement planning.</td>
</tr>
<tr>
<td>American Express</td>
<td>Analyzes data from hundreds of billions of credit card purchases to create “one-to-one” marketing campaigns. Customers receive personalized messages promoting goods and services in which they have shown interest along with their credit card bills.</td>
</tr>
<tr>
<td>U.S. West Communications</td>
<td>Analyzed data from billing operations and external sources to derive customer trends and needs based on household characteristics such as family size, median ages of family members, types of spending patterns, and location. Its findings helped the company increase customer service and reduce the number of lost customers by 45 percent.</td>
</tr>
</tbody>
</table>
Data Mining
Motivation

• Data found in data warehouses is not, by itself, of great intrinsic value.
• Value comes from the knowledge that can be discovered from data.
• What do you do with it?
Data Mining

• Process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

• Extraction of hidden, predictive information from large databases.

• Provide answers to questions a decision maker had previously not thought to ask
Data Mining

• Search for relationships, patterns, and trends which, prior to the search were not known to exist or were not visible.

  E.g. "Find related buying patterns."
  “There is a pattern that occurs X% of the time that when someone buys window coverings (not shades, blinds, or other specifics), and within 1 to 3 months buys linens, within the next 4 months buys furniture.”
Data Mining

Pattern Evaluation

Useful Data

Data Mining

Task-relevant Data

Selection

Data Warehouse

Data Cleansing

Data Preparation

Data Integration

Databases

© 2003 Y. Lee
Examples of Data Mining Analysis Techniques

• Association -- dependency; e.g., buying chips succeeds buying salsa
• Classification -- categories
• Prediction -- suggest outcomes
• Clustering -- data segmentation
• Characterization -- generalization / summarization
Association Rule in Plane Form
Prediction: Categorical Data

Relevance Analysis

Profit:
- \(-365.00\sim 480.00\)
- \(480.00\sim 805.00\)
- \(805.00\sim 1000.00\)
- \(1130.00\sim 1260.00\)
- \(1260.00\sim 6005.00\)

Camping Chain

GO Outlet

Independent

Mass Marketer

Sports Chain

<table>
<thead>
<tr>
<th>Predictive Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sale_Price</td>
<td>(4950.0000 ~ 91950.0000)</td>
</tr>
<tr>
<td>Channel</td>
<td></td>
</tr>
<tr>
<td>Cost_of_Goods_Sold</td>
<td>(-3900.0000 ~ 85900.0000)</td>
</tr>
<tr>
<td>Advertising_Cost</td>
<td>(0.0000 ~ 1715.0000)</td>
</tr>
<tr>
<td>Average_Sales_Area</td>
<td>(1130.0000 ~ 4230.0000)</td>
</tr>
</tbody>
</table>
3D Cube Browser

The image shows a 3D cube browser interface. The browser is used to visualize data across dimensions such as Gender, Customers, Education Level, Graduates, High School Degree, Partial College, High School, and location. The cubes are color-coded and sized to represent different values, such as sales average and profit.
Example Patterns

• Scuba gear and Australian vacations

• Skim milk and whole wheat bread

• AT&T’s stock rises at least 2% after every 3-day slump in DOW
Are “Discovered” Patterns Interesting?

• Data mining query may generate thousands of patterns
• Are they interesting? Why or why not?
• Interesting if:
  – easily understood
  – valid on new or test data with some degree of certainty
  – potentially useful
  – novel
  – validates some hypothesis trying to confirm
Applications: MCI Worldcom

How to find the customers you want to keep from among the millions?

Comb marketing data on 140 million households, each evaluated on as many as 10,000 attributes—e.g. income, lifestyle, and details about past calling habits.

But which set of those attributes is the most important to monitor, and within what range of values?
MCI Worldcom

• Super computer, its data warehouse, has identified variables it finds most telling about it’s customers, and from that, compiled a set of 22 very detailed and highly confidential statistical customer profiles—none of which could have been developed without data mining programs.
Wal-Mart

Point of sale transaction data is captured at each retail store and transmitted to Wal-Mart’s Arkansas data warehouse.

Over 3,500 independent suppliers have online access to information about their respective products in that data warehouse. They may query that data to analyze trends by item and store, using that information to find the products that need replenishment,

and thus allow them to get the right products to each store on time.
Their goal...

track *every item* sold in all its stores in order to find out which in-store location resulted in the *best* sales for a particular product, and to make the *best* merchandising decisions on a *daily* basis.
Wal-Mart

*Triple the size* of data warehouse from *7.5 terabytes* to over *24 terabytes*.

The right item, at the right store, at the right time and price.

“Our business strategy depends on detailed data at every level - every cost, every line item is carefully analyzed enabling better merchandising decisions to be made on a daily basis. It is the foundation for maintaining Wal-mart’s competitive edge and its continuing success in providing everyday low prices and superior customer satisfactions.”

Randy Mott, Senior VP & CIO, Wal-mart
Data Mining Should Not be Used Blindly!

• Data mining find regularities from history
• History not the same as the future
• Association does not dictate trend nor causality
  – Barbecue source, hot dogs and hamburgers
Data Management

COLLECT
- Data Sources
- Represent

STORE
- Databases
- Data Warehouses

MANAGE
- Inhouse vs Outsourcing
- Processing (Transaction versus Web)

USE
- Data Mining
- Target Marketing
Example Use?

• Possible effective use of customer information in many industries
  – Ex. HEB Grocery Case
  – Ex. Two Airline cases
  – Many other mini cases
Conclusion

- Database -- crucial part of traditional and e-commerce business
- Collect and integrate data – internal and external sources
- Database management systems and data warehouses - enable organization of data
- Use – transaction processing, marketing
- Data mining – powerful, strategic tool; results must be carefully assessed
Appendix

• File organization
• Components of database management system
• SQL (Structured Query Language)
FILE ORGANIZATION

• **BIT**: Binary Digit (0,1; Y,N; On,Off)

• **BYTE**: Combination of BITS which represent a CHARACTER

• **FIELD**: Collection of BYTES which represent a DATUM or Fact

• **RECORD**: Collection of FIELDS which reflect a TRANSACTION

Sources: @2002 Prentice Hall
Course Technology, V. Storey
FILE ORGANIZATION

- **FILE**: A Collection of similar RECORDS
- **DATABASE**: An Organization’s Electronic Library of FILES organized to serve business applications

Sources: @2002 Prentice Hall
© 2002 Course Technology
V. Storey
COMPONENTS OF DBMS

• DATA DEFINITION LANGUAGE
  – Defines data elements in database

• DATA MANIPULATION LANGUAGE
  – Manipulates data for applications

• DATA DICTIONARY
  – Formal definitions of all variables in database, controls variety of database contents, data elements

Sources: @2002 Prentice Hall
2003 course Technology
V. Storey
STRUCTURED QUERY LANGUAGE (SQL)

• DE FACTO STANDARD
• DATA MANIPULATION LANGUAGE FOR RELATIONAL DATABASES

Sources: @2002 Prentice Hall
@2002 course Technology
ELEMENTS OF SQL

- **SELECT**: List of columns from tables desired
- **FROM**: Identifies tables from which columns will be selected
- **WHERE**: Includes conditions for selecting specific rows, conditions for joining multiple tables

Sources: @2002 Prentice Hall
@ 2002 Course Technology
V. Storey