Lecture 18

Business Intelligence and Data Warehousing

BDIS 6.2
BSAD 141
Dave Novak

Topics Covered
- Test # Review
- What is Business Intelligence?
- How can an organization be data rich and information poor?
- Terms including Data Warehousing
  - Benefits
  - Multidimensional Analysis
  - Data scrubbing
  - Data mining concepts

Rules for Normalizing "Relationships"

1:1:
- Choose primary key attribute from either entity (but not both) and place copy in the other entity as a Foreign Key.

1:M:
- Place copy of Primary Key attribute of the 1 entity in the M entity as a Foreign Key

M:M:
- Create a joining or intersection table with a foreign key from each of the tables in the M:M relationship
  - Many times the primary key of the M:M Intersection table will be a composite key (a key made up of multiple fields e.g. Student ID and CRN)

1:M Normalization
- Create a Foreign Key on the MANY side of the relationship with the SAME name as the primary key from the 1 side of the relationship

M:M Normalization
- Convert all M:M to two 1:M relationships going to an intersection or joining table

Integrities as Applied to Database Design

Data Integrity:
- Accuracy, Consistency and without Bias.
- Domain Constraints (only #s, currency, only 1 to 10…)
- Field Datatypes on columns (currency, date/time, Text, Boolean…)

Entity Integrity:
- Each Entity has a Unique Identifier which makes every row Unique
- Thus No Duplicate Records
- Each Unique Identifier a.k.a. Primary Key is not null and is in fact Unique
  - All components of the Primary key are NOT null

Referential Integrity:
- All Foreign Key Values EXIST in the Related Primary Key Set
- Also stated as NO orphaned foreign keys
Integrity in MS Access
- Data Integrity – Input Masks and Data Types
  - Zip Codes, Phone Numbers and Date Values
- Entity Integrity – Identifying Field (or Fields)
  - Can’t be Null and auto iterates to a new UNIQUE value
  - Access auto creates an ‘ID’ column set to Auto-Number
- Referential Integrity
  - Can’t put a value in a Foreign Key column that DOESN’T exist in a the Primary Key column to which it relates.

Establishing Relationships in Access
- Referential Integrity requires that ALL values in a foreign key field MUST exist in the Primary Key field of which it is related… other wise it would be an orphaned key related to a record that doesn’t exist.
  - Database are designed to Structure Data
  - Well Designed databases consistently enforce Referential Integrity

Summary
- Test #2 includes definitional concepts as well as practical application of certain concepts
  - Read Chapter 6 and Appendix C
  - Work through all examples
  - Use Cengage online supplements

BDIS 6.2 Main Points
- What is a Data Warehouse (compared to a Database?)
- What is a Data mart (compared to a Data Warehouse?)
- Cluster Analysis and Association Detection
  - Support, Confidence and Lift

Supporting Decisions with Business Intelligence
- Organizational data is difficult to access
- Organizational data contains structured data in database
- Organizational data contains unstructured data such as voice mail, phone calls, text messages, and video clips
The Problem: Data Rich, Information Poor
- Businesses face a data explosion as digital images, email in-boxes, and broadband connections doubles by 2010
- The amount of data generated is doubling every year
- Some believe it will soon double monthly

The Solution: Business Intelligence
- What is BI?
  - Combination of technologies and analytics that facilitate the collection, integration, and presentation of INFORMATION
- Why?
  - Improve decision making

The Solution: Business Intelligence
- Retail and sales: Predicting sales; determining correct inventory levels and distribution schedules among outlets; and loss prevention
- Banking: Forecasting levels of bad loans and fraudulent credit card use, credit card spending by new customers, and which kinds of customers will best respond to (and qualify for) new loan offers

The Solution: Business Intelligence
- Operations management: Predicting machinery failures; finding key factors that control optimization of manufacturing capacity
- Insurance: Forecasting claim amounts and medical coverage costs; classifying the most important elements that affect medical coverage; predicting which customers will buy new insurance policies

The Solution: Business Intelligence
- Improving the quality of business decisions has a direct impact on costs and revenue
- BI enables business users to receive data for analysis that is:
  - Reliable
  - Consistent
  - Understandable
  - Easily manipulated

Term Clarification
- Data Center: Physical space (like a building) to house a large group of servers, storage and network equipment
- Database: A structured set of digitally stored data managed by a computer to support OPERATIONAL needs relating to the creation, retrieval, updates and deletion of data.
- Data Warehouse: a type of database that accumulates and integrates copies of multiple database for ANALYTICAL purposes
- Data Mart: A subset of data from a database or data warehouse used for a single ad-hoc or specific analysis.
Term Clarification

- The definitions imply:
  - Databases are most typically used to support lower-level operational (i.e., largely transactional) decision making
  - Data warehouses integrate data from different DBs and add analytical capabilities needed to support tactical and strategic decision making

Benefits of Data Warehousing

- Data warehouses extend the transformation of data into information
- In the 1990’s executives became less concerned with the day-to-day business operations and more concerned with overall business functions
- The data warehouse provided the ability to support decision making without disrupting the day-to-day operations

Benefits of Data Warehousing

- Data warehouse – A logical collection of information – gathered from many different operational databases – that supports business analysis activities and decision-making tasks
- The primary purpose of a data warehouse is to aggregate information throughout an organization into a single repository for decision-making purposes

Benefits of Data Warehousing

- Inconsistent Data Definitions
  - Every department had its own method for recording data, so when trying to share information, data did not match and users did not get the data they really needed
- Lack of Data Standards
  - Managers need to perform cross-functional analysis using data from all departments, which differed in granularity, format, and levels
- Poor Data Quality
  - The data, if available, were often incorrect or incomplete. Therefore, users could not rely on the data to make decisions

Departmental Databases

- Inadequate Data Usefulness
  - Users could not get the data they needed; what was collected was not always useful for intended purposes
- Ineffective Direct Data Access
  - Most data stored in operational databases did not allow users direct access; users had to wait to have their queries or questions answered by MIS professionals who could code SQL

No Integration Across Departments
**PERFORMING BUSINESS ANALYSIS**

Internal databases
- Marketing
- Sales

External databases
- Competitor information
- Industry information
- Stock market analysis

Data Warehouse Model
- Exploring and mining
- Marketing information
- Inventory information
- Billing information
- Competitor information
- Industry information
- Mailing list information
- Stock market analysis

Databases contain information in a series of two-dimensional tables.

In a data warehouse information is multidimensional, it contains layers of columns and rows.
- **Dimension** – A particular attribute of information
- **Cube** – Common term for the representation of multidimensional information

**Multidimensional Analysis**

**Cubes of Information**

**Benefits of High Quality Information**

- Levels, Formats, and Granularities of Information

**The Cost of Poor Quality Information**

- Butterfly effect
  - Seemingly minor or even insignificant events can have a major impact on a system
- Organizations depend heavily on data
  - A rather simple issue such as misspelling a customer’s name could lead to revenue loss, process problems, and even result in compliance / regulatory issues
The Cost of Poor Quality Information

Information Cleansing or Scrubbing

- An organization must maintain high-quality data in the data warehouse
- Information cleansing or scrubbing – A process that weeds out and fixes or discards inconsistent, incorrect, or incomplete information

Information Cleansing or Scrubbing

- Why is this important in the context of a data warehouse?
  - The data warehouse can contain information from many different DBs – even external DBs
  - The data in these DBs are not consistent with one another

Information Cleansing or Scrubbing

Standardizing Customer Name from Operational Systems

Cost of Accurate and Complete Information
UVM Databases

- Banner – Student Information Systems
  - Courses, Students, Instructors, Grades
- PeopleSoft
  - Salaries, Pay, Tax Information
- BlackBoard
  - Frequency of Logins, Course grades, assignments, readings, materials
- Student Evaluations
  - Each department different – not associable to specific students
- Facilities Swipe Cards
  - Cat Card Swipes for all buildings
- SGA
  - Student Membership/Participation
- UVM Virtual Event Management System
  - Resource Scheduling of all Meeting spaces at UVM
- More……. Alumni, Donors, Research, Meal Plan, Residence halls…

How would UVM use a data warehouse? For example, what types of questions could be addressed that cannot be addressed by using the individual DBs

- Student retention

Trends and Patterns with Data Mining

- Common forms of data-mining analysis capabilities include
  - Cluster analysis
  - Association detection
  - Statistical analysis

Cluster Analysis

- Statistical approach used to classify data or objects (people, places, or things) into distinct predefined groups
- The objects within a particular group are very similar to one another and are very different from the objects in other groups
- Provides insight into association and pattern

Cluster Analysis

- Marketing – group all customers into distinct groups based on their purchasing patterns
- Insurance – identify automobile policy holders with a higher than average claim cost
- Finance – identify certain categories of stocks that exceed some set of pre-determined performance thresholds

Association Detection = Classic Example being ‘Market Basket Analysis’

- Market-basket analysis is a data-mining technique for determining sales patterns
  - Uses statistical methods to identify sales patterns in large volumes of data
  - Shows which products customers tend to buy together
  - Used to estimate probability of customer purchase
  - Helps identify cross-selling opportunities
- “Customers who bought book X also bought book Y”
Market Basket Analysis
- Information Systems for Large Retail Organizations Creating Mounds of Data
- Transactional Data about:
  - What was bought
  - What time of day
  - What day of week
  - What time of year
  - AND what else was bought that Visit
- Incomplete demographic information about the customer

Market Basket Analysis
- Categorize all products into a finite, manageable number of categories
  - Beer
  - Vegetables
  - Deli
  - Soda
  - Frozen dinners…
- Identify Relationships Between Product Categories

Market Basket Analysis
- Results of the Data mining:
  - Unexpected Yet Logical
  - So what do you with this information?

Market Basket Analysis
- Sample of 1,000 receipts
  - 100 of them had BOTH diapers and beer
    - \(0.10\) Support (100/1000)
  - 125 had Diapers (but not necessarily beer)
    - \(0.80\) Confidence (125/150)
  - 200 of them bought Beer (but not necessarily diapers)
    - \(0.20\) Base Probability (200/1000)
- Lift = \(\frac{\text{Confidence of Diapers & Beer}}{\text{Base Probability of Beer}}\)
  - Lift shows 4 times more likely to buy beer when they buy diapers (0.80/0.20)
- [http://www.uvm.edu/~tichitte/mba/mba.html](http://www.uvm.edu/~tichitte/mba/mba.html)

Summary
- Test # Review
- What is Business Intelligence?
- How can an organization be data rich and information poor?
- Terms including Data Warehousing
  - Benefits
- Multidimensional Analysis
- Data scrubbing
- Data mining concepts