Data Warehousing & Data Mining

Course Introduction
Objectives & Outline
Expectations

Lecture 1 Dated 18/10/2010

Dr. Maruf Pasha
Computer Science Department
Bahauddin Zakariya University

Data Warehousing and Data Mining

- Instructor:
  - Maruf Pasha
    Email: maruf_pasha@yahoo.com

2

Objectives

- Provide the students with a complete background on data warehousing and data mining, basic algorithms, essential concepts and popular techniques.
- Introduce the latest works in specific domains within data warehousing data mining and involve the students in reading and critical analysis of reported literature from authentic sources.
- Equip the students with sufficient literature knowledge so that future projects may be identified.

3

Pre-Reqs

- **Essential:** Undergrad courses in Algorithms and Data Structures, Mathematics, Databases, Object Oriented Programming (C/C++/JAVA)
- **Beneficial:** Undergrad courses in Linear Algebra, Data Warehousing.

4

Text and Reference Material

- The course will be mainly based on research literature, following text may however be consulted:
  - Principles of Database and Knowledge-base Systems – Volume 1, by Jeffrey Ullman or any other basic database textbook (chapters 2, 4, 7)
  - Database System Concepts, by A. Silberschatz, H.F.Korth, S. Sudarshan (chapter 7)
  - Ralph Kimball, Joe Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data
  - H.W. Inmon, The Data Warehouse Environment: Building the Data Warehouse
  - Ralph Kimball, Margy Ross, Data Warehouse Toolkit: The complete Guide to Dimensional Modeling

5

Text and Reference Material

- David Hand, Heikki Mannila and Padhraic Smyth. “Principles of Data Mining”. Pub. Prentice Hall of India, 2004.
- Sushmita Mitra and Tinku Acharya. “Data Mining: Multimedia, Soft Computing and Bioinformatics”. Pub. Wiley an Sons Inc. 2003
- Usama M. Fayyad et al. “Advances in Knowledge Discovery and Data Mining”, The MIT Press, 1996.

6

Grading Criteria (tentative)

Item	Weight (%)
Quiz	5
Assignments	5
Term Project	8
Attendance	2
Midterm	30
Final	50

7

Course Outline

DATA WAREHOUSE OVERVIEW

- Overview
- Typical uses

DEFINITION, ARCHITECTURE AND CONCEPTS

- Enterprise Data Model
- Operational vs. historical data
- Extract Transform Load (ETL)
- Metadata
- Data warehouse vs. data mart
- Data mining
- OLAP vs. OLTP
- Massive size implementation
- Logical design vs. physical design
- Normalization vs. denormalization
- Referential constraints

8

Course Outline

DATA MODELLING OPTIONS

- -Entity model
- -Star schema
- -Snowflake schema

DIMENSIONAL MODELLING DESIGN

- -Overview
- -Metadata properties
- -Star schema
- -Snowflake schema
- -Cubes
- -Measures and facts
- -Attributes and relationships
- -Dimension
- -Hierarchies
- -Joins
- -Summary tables and aggregation

9

Course Outline

OLTP VS. OLAP APPLICATION
ETL
DATA MINING
-ALGORTIHMS
AND LIST GOES ON……..

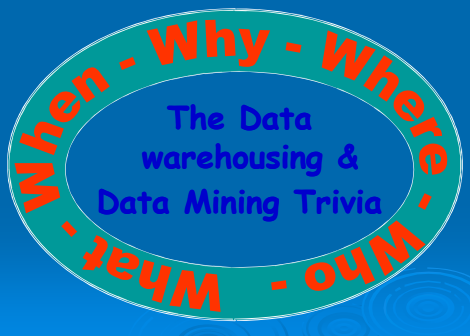
10

Expectations

- Remember.
  - Quality, Extent and Depth of contents largely depends on class attitude.
  - Take-up the lead yourself.
- Advice on deliverables
  - Start early
  - Discuss progress regularly
  - Make an appropriate group
  - Honesty is not the best policy, it's the **Only** policy
  - Zero Credit for copying/plagiarism/late submission.

11

Data Warehousing & Data Mining –



12

Background

- 1980's to early 1990's
  - Focus on computerizing business processes
  - To gain competitive advantage
- By early 1990's
  - All companies had operational systems
  - It no longer offered any advantage
- How to get competitive advantage??

Need for Data Warehouse

- Companies, over the years, gathered huge volumes of data
- "Hidden Treasure"
- Can this data be used in any way?
- Can we analyze this data to get any competitive advantage?

Data Warehousing

- Abbreviated *DW*, a collection of data designed to support management decision making. Data warehouses contain a wide variety of data that present a coherent picture of business conditions at a single point in time.
- Development of a data warehouse includes development of systems to extract data from operating systems plus installation of a warehouse database system that provides managers flexible access to the data.
- The term data warehousing generally refers to the combination of many different databases across an entire enterprise. Contrast with *data mart*

15

Benefits of Data Warehousing

- Allows "efficient" analysis of data
- Competitive Advantage
- Analysis aids strategic decision making
- Increased productivity of decision makers
- Potential high ROI

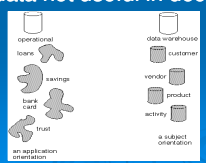
Inmons's definition

A data warehouse is
-subject-oriented,
-integrated,
-time-variant,
-nonvolatile
collection of data in support of management's decision making process.

17

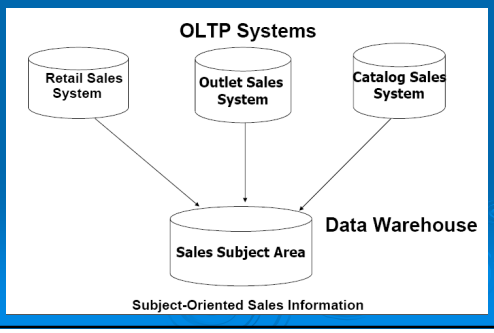
Subject-oriented

- Data warehouse is organized around subjects such as sales,product,customer.
- It focuses on modeling and analysis of data for decision makers.
- Excludes data not useful in decision support process.



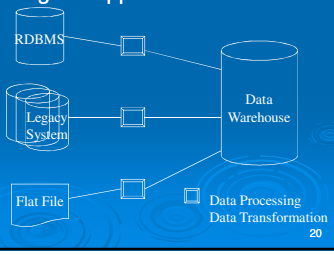
18

Subject-oriented



Integration

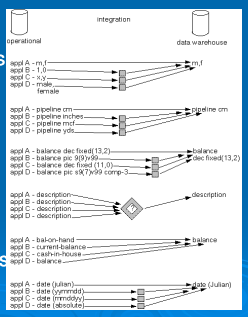
- Data Warehouse is constructed by integrating multiple heterogeneous sources.
- Data Preprocessing are applied to ensure consistency.



20

Integration

- In terms of data.
  - encoding structures
  - Measurement of attributes.
  - physical attribute. of data
  - naming conventions
  - Data type format



21

Time-variant

- Provides information from historical perspective e.g. past 5-10 years
- Every key structure contains either implicitly or explicitly an element of time

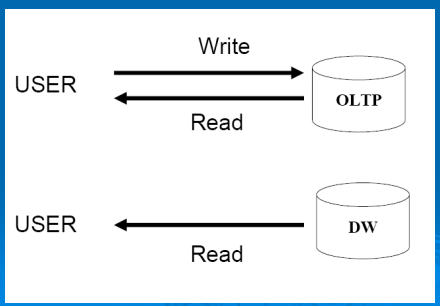
22

Nonvolatile

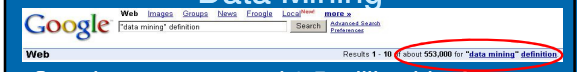
- Data once recorded cannot be updated.
- Data warehouse requires two operations in data accessing
  - Initial loading of data
  - Access of data

23

Nonvolatile



Data Mining



- Google returns around 0.5 million hits for "Data Mining" + Definition
- Some common terms shared by almost all definitions
  - Discovery – Hidden Patterns – Non trivial – Process
- What's in a Name
  - Information Extraction, Information Retrieval (IR), Knowledge Discovery, Knowledge Discovery in Databases (KDD), Pattern Detection etc.

25

Intelligent Problem Solving

- Knowledge = Facts + Beliefs + Heuristics
- Success = Finding a good-enough answer with the resources available
- Search efficiency directly affects success

26

Focus on Knowledge

- Several difficult problems do not have tractable algorithmic solutions
- Human experts achieve high level of performance through the application of quality knowledge
- Knowledge in itself is a resource. Extracting it from humans and putting it in computable forms reduces the cost of knowledge reproduction and exploitation

27

Value of Information

- Exponential growth in information storage
- Tremendous increase in information retrieval
- Information is a factor of production
- Knowledge is lost due to information overload

28

KDD vs. DM

- Knowledge discovery in databases
  - "non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data"
- Data mining
  - Discovery stage of KDD

29

Questions

30