CSPP 53017: Data Warehousing

Winter 2013

Lecture 1

Svetlozar Nestorov

Class Organization

- Lectures
- Mailing list
- Office hours
- Homework
- Team Projects
- Team Presentations
- Quizzes
- Grades

Class Details

- Recommended books:
  - The Data Warehouse Toolkit (2nd edition)
  - The Data Warehouse Lifecycle Toolkit (2nd edition)
  - The Data Warehouse ETL Toolkit
  - all three books by Kimball and Ross
  - Building the Data Warehouse (4th edition) by Inmon
- Gradiance
  - Online homework and assessment system.
- Emails
  - Please, use **csppdw** in the subject of your email.

Lectures

- Discussion oriented
  - Ask questions!
- Mix of slides and whiteboard examples
  - Not everything we cover will be on the slides!
- Team presentations in class
- Three 15-minute in-class quizzes
- Every Tuesday evening from 5:30pm to 8:30pm in Gleacher 604 until March 12.

Course Work

- Occasional online homework
  - Solve and submit online on Gradiance.
  - Class token: **EFEE5549**
- Weekly multipart project:
  - Design and develop a (limited) real world data warehouse
- Exams: 3 in-class 15-minute quizzes

Class Overview

- Data warehousing history and motivation
- Basic elements and processes
- The Lifecycle approach
- Data design: dimensional modeling and the star schema
- On-Line Analytical Processing (OLAP)
- ETL design and development
- Metadata
- DW growth and development
- Data mining

Suggested Projects: Datasets

- Nielsen datasets
  - Homescan data
  - Media data
- City of Chicago data
  - crime reports
  - 311calls
  - building permits
  - business licenses
  - bike racks
  - …
- Your project?

Class Introductions

- Name
- General background
  - Techie, manager, poet, adventurer?
- Database and data warehousing experience
  - Systems used
  - Project sizes (data, team, code)
- Class interests and goals
  - General interest?
  - Relevance to current job?
  - Specific project?
- Anything else: suggest advanced topics to discuss?

Initial Terminology

- DBMS - Data Base Management Systems
- DSS - Decision Support Systems
  - systems that facilitate data processing used to drive management decisions (as opposed to transaction processing exclusively used to drive detailed operational decisions)

Evolution of Decision Support Systems

1970's

- DASD
  - Direct Access Storage Devices
- DBMS
  - Data Base Management Systems
- OLTP – On Line Transaction Processing
  - A type of computer processing in which the computer responds immediately to user requests
- OLTP on Databases
- Database defined as a single source of data for all processing

Evolution of Decision Support Systems

1980's

- PC/4GL Technology
  - end-users able to directly control data and systems (notion that data can be used for more than operational transactional purposes)
- Management Information System
  - MIS (later known as Decision Support Systems-DSS) defined as processing used to drive management decisions (as opposed to processing exclusively used to drive detailed operational decisions)
- Still *single-database-serving-all-purpose* paradigm
  - One DBMS supporting both regular transaction processing and MIS/DSS
- Extract Processing
  - Extract Program
    - rummages through a file or database, uses some criteria for selection, and upon finding qualified data, transports the data over onto another file or database
    - used to move data used for DSS out of the way of high performance on line transaction processing and to give the DSS user the control over that data

Evolution of Decision Support Systems

1990's

- Proliferation of Extracts
  - Extracts everywhere (and extracts of extracts, and extracts of extracts of extracts, and …)
- Productivity Problems
  - In order to write a corporate report
    - Data must be located
    - Lots of customized programs must be written
    - The programs must cross every technology that the company has
- Often, inability to go from Data to Information
  - Due to the following:
    - Data Applications were built to serve the needs of current transaction processing
    - They were never designed to hold the historical data needed for DSS analysis

Evolution of Decision Support Systems

Early 1990's

- Lack of Credibility of Data

E.g. Market share activity report done by 2 analysts

- No common time basis for data (e.g. one analyst extracts data on Monday morning, another extracts data on Wednesday afternoon)
- Algorithmic differential (e.g. one analyst extracts data on all sales, another extracts data on all food related sales)
- The levels of extraction - each additional level of extraction increases the probability of discrepancy (e.g. one analyst uses an extract, another uses an extract of an extract)
- External data (e.g. one analyst is bringing in Wall Street Journal data, another is bringing in Business Week data, and they both strip the data identity)
- No common source of data (e.g. analysts are extracting data from different databases within the company)

Final Result:

Analyst A	Analyst B
Activity is up 10%	***Activity is down 15%***

Evolution of Decision Support Systems

- Example – Accenture Survey*
  - Accenture surveyed 1,000 managers at large enterprises and found that managers spend up to two hours a day searching for information, and more than 50 percent of the information they obtain has no value to them.
  - Only half of all managers believe their companies do a good job in governing information distribution or have established adequate processes to determine what data each part of an organization needs.
  - 59 percent said that as a consequence of poor information distribution, they miss information that might be valuable to their jobs almost every day because it exists somewhere else in the company and they just can not find it.
  - 42 percent of respondents said they accidentally use the wrong information at least once a week.
  - The outcome: the terabytes of data that enterprises gather -- and spend millions on storing, managing and analyzing -- is bordering on useless for decision-makers. The cause for this sad state of affairs is the usual culprit: difficult to access, poorly-integrated and siloed data.

*Jan 4 2007 :

Evolution of Decision Support Systems

- Approach Change
  - Realization that Naturally Evolving Architecture (direct result of extract processing method) is not sufficient
- Architected Environment
  - Recognition that there are fundamentally two kinds of data
    - OPERATIONAL DATA
    - ANALYTICAL DATA
  - Data Warehouse emerged as the New DSS Architecture

Need for Data Warehousing

- Integrated, company-wide view of high-quality information.
- Separation of **operational** and **analytical** systems and data.

OPERATIONAL vs. ANALYTICAL DATA

Operational Data	Analytical Data
Data Differences	
Typical Time-Horizon: Days/Months	Typical Time-Horizon: Years
Detailed	Summarized (and/or Detailed)
Current	Values over time (Snapshots)
Technical Differences	
Can be Updated	Read (and Append) Only
Control of Update: Major Issue	Control of Update: No Issue
Small Amounts used in a Process	Large Amounts used in a Process
Non-Redundant	Redundancy not an Issue
High frequency of Access	Low/Modest frequency of Access
Purpose Differences	
Supports Day-to-Day Operations	Supports Managerial Needs
Application Oriented	Subject Oriented

Application vs. Subject Oriented

Application:
Health Club Members-Visit Database

HEALTHCLUBMEMBERS

MembId	Name	MembLevel	DatePayed
111	Joe	A	01/01/2008
222	Sue	B	01/01/2008
333	Pat	A	01/01/2008
…	…	…	…

DAILYVISITSFROMNONMEMBERS

Trid	VisitType	VisitDate
11xx22	YP	01/01/2008
11xx23	NP	02/01/2008
11xx24	YP	02/01/2008
…	…	…

MEMBRSHPLEVELS

ID	Type	Fee
A	Gold	$100
B	Basic	$50

VISITLEVELS

ID	Type	Fee
YP	With Pool Usage	$15
NP	Without Pool Usage	$10

Application vs. Subject Oriented

Application:
Health Club Members-Visit Database

HEALTHCLUBMEMBERS

MembId	Name	MembLevel	DatePayed
111	Joe	A	01/01/2008
222	Sue	B	01/01/2008
333	Pat	A	01/01/2008
…	…	…	…

DAILYVISITSFROMNONMEMBERS

Trid	VisitType	VisitDate
11xx22	YP	01/01/2008
11xx23	NP	02/01/2008
11xx24	YP	02/01/2008
…	…	…

MEMBRSHPLEVELS

ID	Type	Fee
A	Gold	$100
B	Basic	$50

VISITLEVELS

ID	Type	Fee
YP	With Pool Usage	$15
NP	Without Pool Usage	$10

Subject:
Health Club Revenue

REVENUE

Rid	Date	GeneratedBy	Amount
7235	01/01/2008	NonMember	$15
7236	01/01/2008	Member	$100
7237	01/01/2008	Member	$50
7238	01/01/2008	Member	$100
7239	02/01/2008	NonMember	$10
7240	02/01/2008	NonMember	$15
…	…	…	…

Application vs. Subject Oriented

Application:
Health Club Members-Visit Database

HEALTHCLUBMEMEBRS

MembId	Name	MembLevel	DatePayed
111	Joe	A	01/01/2008
222	Sue	B	01/01/2008
333	Pat	A	01/01/2008
…	…	…	…

DAILYVISITSFROMNONMEMBERS

Trid	VisitType	VisitDate
11xx22	YP	01/01/2008
11xx23	NP	02/01/2008
11xx24	YP	02/01/2008
…	…	…

MEMBRSHPLEVELS

ID	Type	Fee
A	Gold	$100
B	Basic	$50

VISITLEVELS

ID	Type	Fee
YP	With Pool Usage	$15
NP	Without Pool Usage	$10

Subject:
Health Club Revenue

REVENUE

Rid	Date	GeneratedBy	Amount
7235	01/01/2008	NonMember	$15
7236	01/01/2008	Member	$100
7237	01/01/2008	Member	$50
7238	01/01/2008	Member	$100
7239	02/01/2008	NonMember	$10
7240	02/01/2008	NonMember	$15
…	…	…	…

Application vs. Subject Oriented

Application:
Health Club Members-Visit Database

HEALTHCLUBMEMEBRS

MembId	Name	MembLevel	DatePayed
111	Joe	A	01/01/2008
222	Sue	B	01/01/2008
333	Pat	A	01/01/2008
…	…	…	…

DAILYVISITSFROMNONMEMBERS

Trid	VisitType	VisitDate
11xx22	YP	01/01/2008
11xx23	**NP**	**02/01/2008**
11xx24	YP	02/01/2008
…	…	…

MEMBRSHPLEVELS

ID	Type	Fee
A	Gold	$100
B	Basic	$50

VISITLEVELS

ID	Type	Fee
YP	With Pool Usage	$15
NP	**Without Pool Usage**	**$10**

Subject:
Health Club Revenue

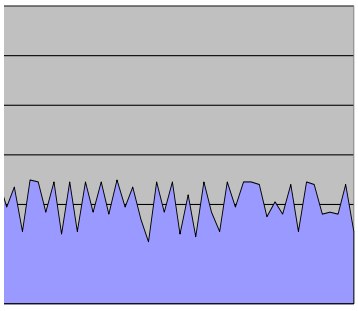
REVENUE

Rid	Date	GeneratedBy	Amount
7235	01/01/2008	NonMember	$15
7236	01/01/2008	Member	$100
7237	01/01/2008	Member	$50
7238	01/01/2008	Member	$100
7239	**02/01/2008**	**NonMember**	**$10**
7240	02/01/2008	NonMember	$15
…	…	…	…

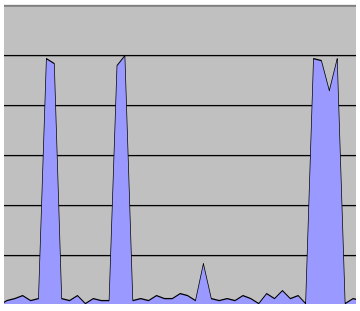
OPERATIONAL vs. ANALYTICAL DATA

Hardware Utilization
(*Frequency and Amount of Data Access*)

Operational



Data Warehouse



Data Warehouse: Definition

- Data Warehouse: An enterprise-wide structured repository of subject-oriented, time-variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data.

Bill Inmon
(paraphrased by Oracle Data Warehouse Method)

Data Warehouse: Definition

- Data Warehouse: An **enterprise-wide** structured repository of subject-oriented, time-variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data.

- **enterprise-wide** refers to the fact that a DWH provides a company-wide view of the information it contains

Data Warehouse: Definition

- Data Warehouse: An enterprise-wide **structured repository** of subject-oriented, time-variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data.

- **structured repository** refers to the fact that a DWH is a structured data repository like any other database

Data Warehouse: Definition

- Data Warehouse: An enterprise-wide structured repository of **subject-oriented**, time-variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data.

- **subject-oriented** refers to the fundamental difference in the purpose of a traditional Database System and a DWH.
  - Traditional Database System is developed in order to support a specific business operation (e.g. shipping company order-entry database, dental office appointment management database).
  - DWH is developed to analyze a specific business subject area (e.g. sales, profit).

Data Warehouse: Definition

- Data Warehouse: An enterprise-wide structured repository of subject-oriented, **time-variant**, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data.

- **time-variant** refers to the fact that a DWH contains slices of data across different periods of time. With these data slices, the user can view reports based on current as well as past data.

Data Warehouse: Definition

- Data Warehouse: An enterprise-wide structured repository of subject-oriented, time-variant, **historical** data used for information retrieval and decision support. The data warehouse stores atomic and summary data.

- **historical** refers to the fact that a DWH typically contains several years worth of data (as opposed to the typical 60-90 days time horizon for data in many traditional operational databases).

Data Warehouse: Definition

- Data Warehouse: An enterprise-wide structured repository of subject-oriented, time-variant, historical data used for **information retrieval and decision support**. The data warehouse stores atomic and summary data.

- **information retrieval and decision support** refers to the fact that a DWH is a facility for getting information to answer questions (it is not meant for direct data entry; batch updates are the norm for refreshing data warehouses) of analytical and strategic nature

Data Warehouse: Definition

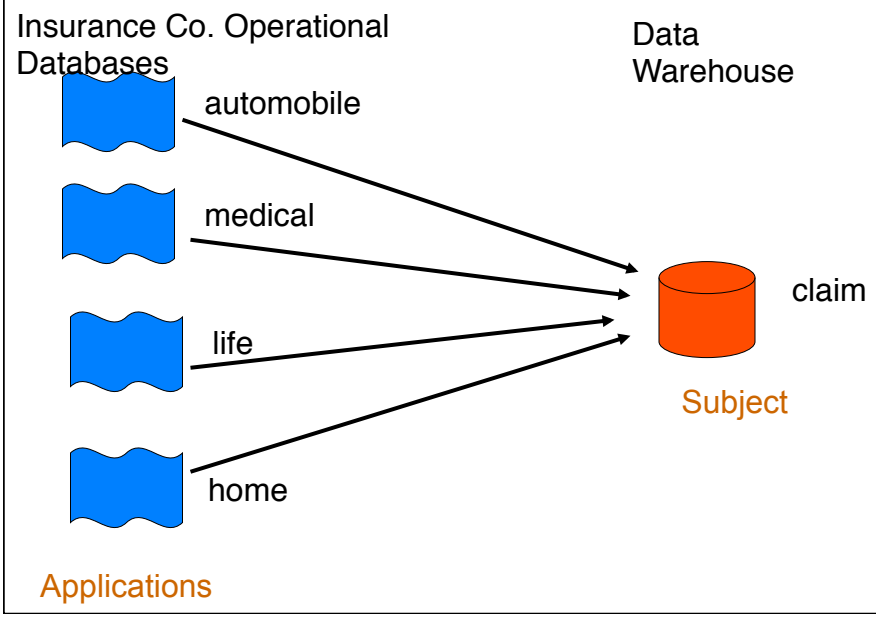
- Data Warehouse: An enterprise-wide structured repository of subject-oriented, time-variant, historical data used for information retrieval and decision support. The data warehouse stores **atomic and summary data**.

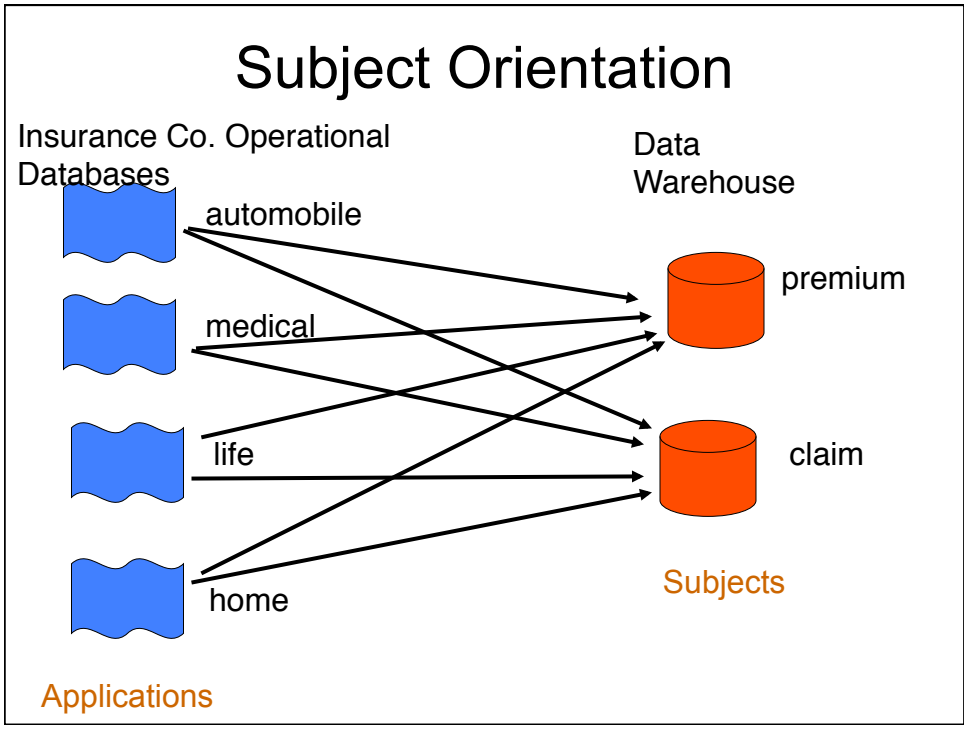
- **atomic and summary data** refers to the fact that a DWH, depending on purpose, may contain atomic data, summary data, or both.

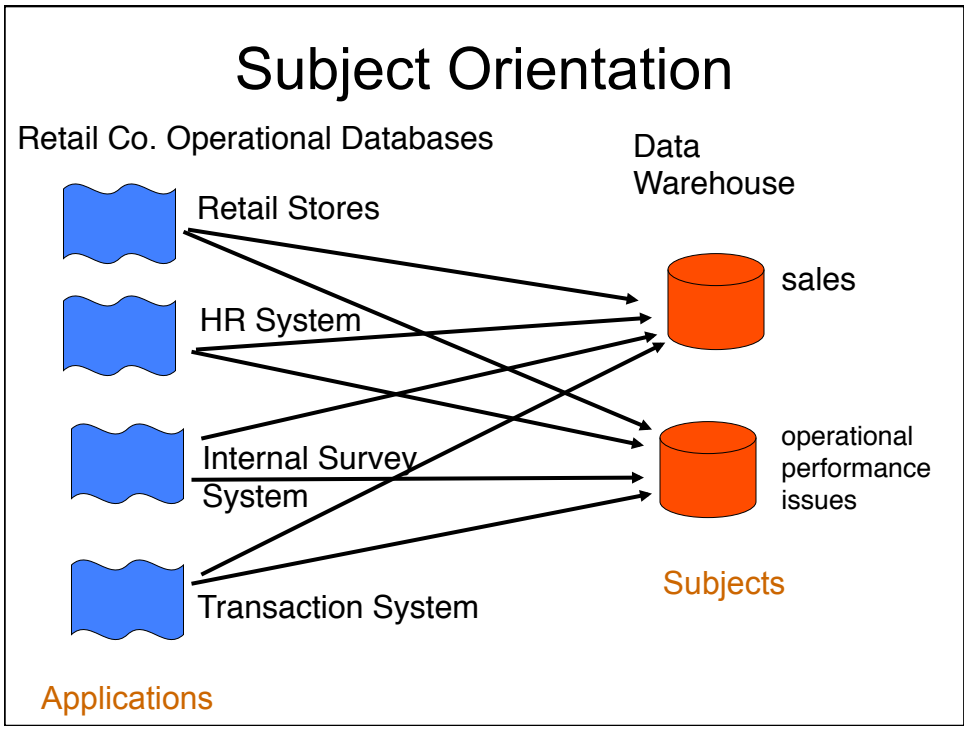
DWH is Subject Oriented!

- Data is organized around major subject areas of an enterprise, and is therefore useful for an enterprise-wide understanding of those subjects
- Data from operational systems must be transformed so that is consistent and meaningful in the DWH

Subject Orientation




Subject Orientation
Insurance Co. Operational Databases
Data Warehouse
automobile
medical
life
home
premium
claim
Subjects
Applications


Subject Orientation
Retail Co. Operational Databases
Data Warehouse
Retail Stores
HR System
Internal Survey System
Transaction System
sales
operational performance issues
Subjects
Applications

DW is Integrated!

- In many organizations, data resides in diverse independent systems, making it difficult to acquire meaningful information for analysis.
- In DWH data is completely integrated, even when the underlying sources store data differently
- Unfortunately, there is no magic wand
  - Instead we have the transformation and integration process (which involves **ETL** – extraction, transformation, and load)
  - Building the **ETL** infrastructure and using it to move data from source systems into a data warehouse (data staging) can be time consuming and costly
  - In many cases, the majority of time within a DWH project is spent on the data staging phase (building and utilizing **ETL** infrastructure)

Common Example of a Data Warehouse Purpose

- Data warehouse is often designed and implemented to answer TWO fundamental questions:
  - Who is buying what?
  - When and where are they doing so?
- More specific
  - Who [which customer] is buying [buying / using / delivering / shipping / ordering / returning] what [products / services] from where [outlet / store / clinic / branch] on what occasion [when], how [credit card / cash / check / exchange / debit] and why [causation]?

Some Uses of a Data Warehouse

- Airlines for aircraft deployment, analysis of route profitability, frequent flyer promotions, and maintenance
- Banks for promotion of products and services, and customer service
- Health care for cost reduction
- Investment and insurance companies for customer analysis, risk assessment, and portfolio management
- Retail stores for buying pattern analysis, promotions, customer profiling, and pricing
- Telecommunications for product and service promotions.