Intro to Data Warehousing

What is a data warehouse?
A data warehouse is an integrated, subject-oriented, time-variant, non-volatile database that provides support for decision making.

Bill Inmon
(The “Father” of data warehousing)

What is meant by each of the underlined words in this definition?

Evolution of Data Warehousing

Business drivers:
- strategic value of operational data
- need for rapid decision-making
- need to push decision-making to lower levels in the organization

The business issue:
Transformation of ‘data’ into ‘information’ for decision-making

Late 1990s
Sophisticated Data Warehouse Capabilities

Mid-1990s
Early Data Warehouses

Early 1990s
‘Standalone’ DSSs

Early 1980s
Ad Hoc Query Tools
(RDBMS, SQL)

Mid-1980s
Early Reporting Systems

Early 1980s
Early Reporting Systems
Example Data Warehouse Projects

- **General Motors**
  - building a central customer data warehouse
  - will hold customer data from all car and truck
  - divisions, car leasing, and home mortgage and credit-card units
  - enable them to ‘cross-sell’ customers on GM financial
  - services and target marketing (e.g., offer Information
  - on minivans to a customer who has had a baby)
  - Source: Computerworld, 7/5/99

- **Anthem Blue Cross and Blue Shield**
  - created a common repository for claims, revenue, and
  - services provided by hospitals and physicians
  - patient benefits: identified certain providers with
  - superior success rates for coronary artery bypass
  - surgery (reduced mortality rate from more than 4% to
  - less than 1% for policyholders)
  - Source: CACM, 9/98

Business Benefits of Data Warehousing

- **Time savings**
- **More and better information**
- **Better decisions**
- **Support for accomplishing strategic business objectives**

Source: Watson and Haley, CACM, 9/98

Operational vs. Data Warehouse Data

- **Operational data is organized around functional
  organizations within a business**
- **Data warehouse data is subject-oriented - with
  an enterprise view**
- **Operational data and data warehouse data differ in
  three main areas:**
  - **Timespan**
    - Operational data: represents current transactions
      (e.g., General Ledger data as of 9/10/04)
    - Data warehouse data: covers longer timeframes
      (e.g., General Ledger data for the past 5 years)
Operational vs. Data Warehouse Data

Operational data and data warehouse data differ in three main areas (cont'd):

- **Granularity**
  - **Operational data**: specific transactions that occur at a given time (e.g., amount of the last salary increase for employee 992)
  - **Data warehouse data**: data presented at different levels of aggregation (e.g., average salary increase for engineers, managers)

- **Dimensionality**
  - **Operational data**: one-dimensional view of the data (atomic transactions - e.g., payment by customer X)
  - **Data warehouse data**: multi-dimensional view of the data (e.g., how many widgets of type X were sold to customer Y during the last six months?)

Basic Components of a Data Warehouse

- **Data extraction, filtering, and loading component**
  - software that extracts, validates and transforms the data taken from the operational databases (e.g., HR/Payroll database) and the external data sources
  - software that loads the data into the data warehouse database

- **Data storage component**
  - data warehouse database

- **Metadata component**
  - contains 'data about the data' - data schemas, business rules, etc.

Basic Components of a Data Warehouse

- **“Canned” reports and screens**
  - standard end-user reports and screens

- **Ad hoc report tools**
  - end-user tools to create custom reports as needed

- **Sophisticated end-user analytical tools**
  - tools for data analysis, data modeling and presentation, etc. Ex: On-Line Analytical Processing tools (OLAP)
**Data Mining**

- Data mining tools automatically search data for anomalies and possible relationships
- analyze the data
- uncover problems or opportunities hidden in the data
- form computer models based on their findings
- predict business behavior using the models

- Operational data, data warehouse data, and external data can all be 'mined'

**Data Warehousing - Trends & Issues**

- **Pervasiveness:** Most organizations now have some type of a data warehouse
- **Web access:** Providing access to data warehouses over the Web
- **Shared data warehouses:** Opening data warehouses to customers, suppliers and other business partners
  - e.g., GM's Supply Chain Data Warehouse
    - has information about quantities of supplies shipped, delivery times, and prices
    - will be available via the Web to more than 5,000 suppliers and supplier organization members