Data Warehouses for Integration

- What are data warehouses
- Why they are needed for integration
- Differences between operational and decision support systems
- Data warehouse architectures
- Web Data Warehousing
- Meta data and Data Modeling
- Dimensional Data Modeling
- Data mining overview
- Data Mining
- Web Mining

Amjad Umar

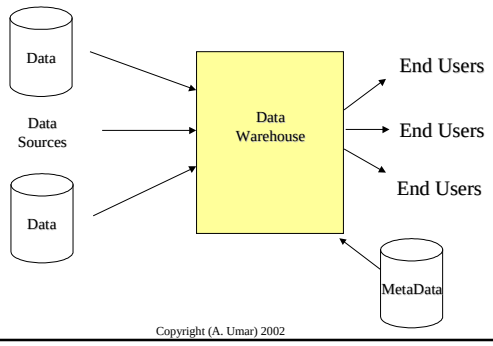
Data Warehouse

Data warehouse: Information repository for decision support

Data Warehouse characteristics:

- Contains summary (light and heavy) and historical data
- Data may be from operational systems, external systems, unstructured data (memos, letters, pictures)
- Intended for decision support (primarily retrieval, discovery) and not by applications (order processing, purchasing and sales)
- Data is subject-oriented (customers, products)
- Support for tools:
  - Report writers
  - Data browsers
  - Spreadsheets
  - 4GL (e.g., focus)
  - "Drill down" applications
  - Data mining tools
  - Other planning and modeling tools

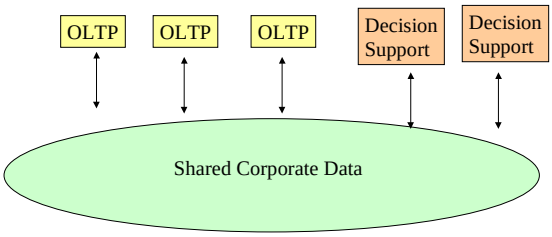
Data Warehousing Logical Architecture



Sample Queries (actual)

- “How many customers have complained more than once for our new software”
- How many customers had to call the customer contact center more than once for the same problem”
- “How many people have claimed loss luggage on our airline”
- "Which products in my store are selling most quickly?"
- "Which products stay in the inventory the longest?"
- "How does one store's profitability compare to the rest of the chain?"
- "How many VISA card holders from our bank did not use their VISA card last year".The answer was: 500,000. This meant that the bank paid $12 million dollars to VISA for nothing
- "Which doctors in California charge more than the national average for a broken leg procedure?"
- "Which doctors filed the largest number of claims during the last quarter?"
- "How many purchase orders were placed in 1991, 1992, and 1993 for the Sector, and what were the corresponding dollars".
- "What are the top 10 purchased commodities and corresponding suppliers in the Sector"

Can these queries be satisfied by a common corporate data?



What type of technologies will make it possible?

Why Data Warehouses for Integration

- Data of legacy apps can be extracted and loaded in a DW in stead of re-engg the legacy app
- DW avoids interference with operations (queries may interfere with the day-to-day operation of the on-line transaction processing applications).
- DW can be used for data spread across multiple systems stored in diverse database managers requiring diverse access methods (may require expensive mediators).
- DW can be used for legacy data that is embedded in IMS databases and flat files that do not support adhoc queries
- DW can be used for more indexes for extensive queries
- Others?

Data Warehouse: To do or not to do

Data warehouse is, in general, a good approach if:

- Demand for ad hoc queries and analysis is very high
- The needed data is used for decision support (it is easier to provide decision support through a data warehouse)
- The surround technology for access in place is not efficient and does not meet the requirements of new users
- The data does not change frequently (in such cases, data warehouse needs to be synchronized with the back-end system frequently)
- Needed data is embedded n too many legacy systems (it is difficult to directly access 30 or so data sources and perform joins among them, etc.)

Data warehouses are not a good choice if:

- Demand for data is low (access in place is better in this case)
- Most recent copy of data is essential (DWs give somewhat outdated view)
- Data changes frequently in the sources

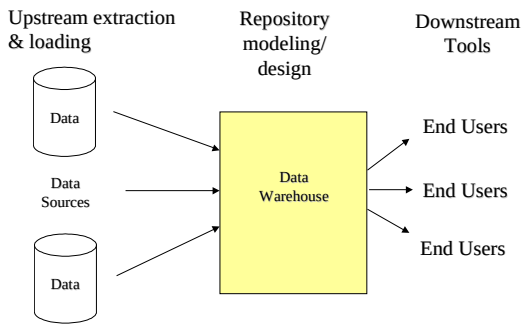
In general, data warehouses are useful in their own right because they support the decision makers in organizations.

However, they create duplicate data

Data Warehouse Architecture Overview

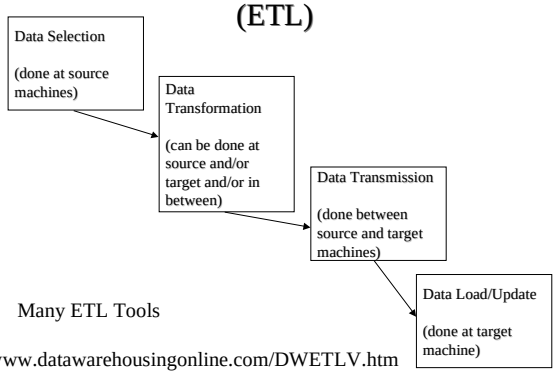
- Common Mechanism:
  - Extract information from operational data
  - Store results of reports in the database
  - Design the DW DB (Highly or lightly summarize data)
  - Synchronize data periodically
- Data mirroring is a special case

Data Warehouse Design Considerations



Many choices and tradeoffs at each step

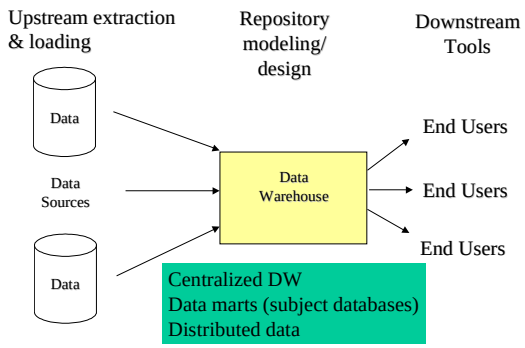
Data Extraction/Transformation/Load (ETL)



Many ETL Tools

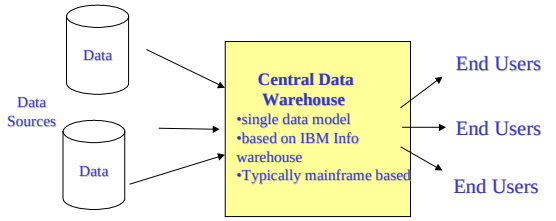
www.datawarehousingonline.com/DWETLV.htm

Repository Design



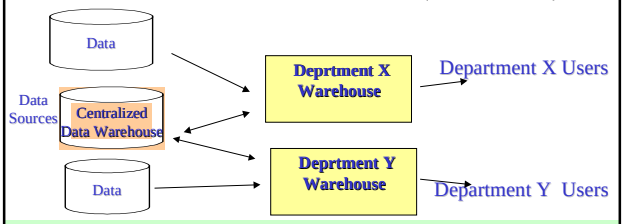
Centralized DW
Data marts (subject databases)
Distributed data

Centralized Data Warehouse



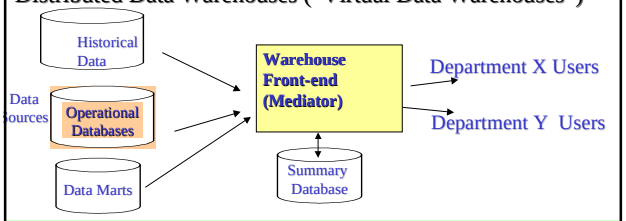
Plusses:
•same consistent and complete data for all users.
•users logon to one environment (no multiple data sources)
•Works well where most of the processing is done at the corporate
Minusses:
•very difficult to develop a global data model for most organizations
•difficult to agree on a corporate wide level of detail and naming conventions.
•need to carefully manage the performance and end-user access

Local Functional Warehouses (Data Marts)



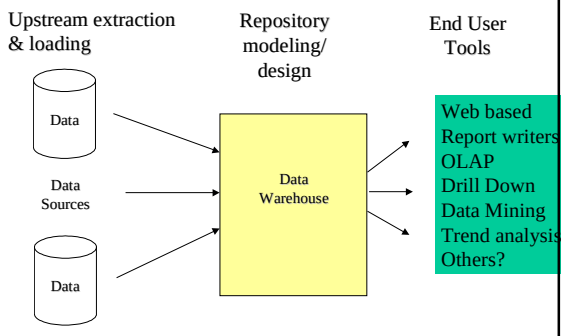
- Typically created by departments/divisions to support their own decisions
- May be created to support specific products (e.g., automobile parts) or function (e.g., loan management)
- May be created for user populations/environments (DW for PC users)
- May be fed by a centralized DW
- **Plus**: can be developed quickly to serve the local needs without having to wait for the large corporate data warehouse.
- **Minus**: proliferation of DWs that are not consistent with each other.
- **In practice**, data marts can be used by independent departments as a starting point in an overall strategy for a centralized corporate data warehouse.

Distributed Data Warehouses ("Virtual Data Warehouses")

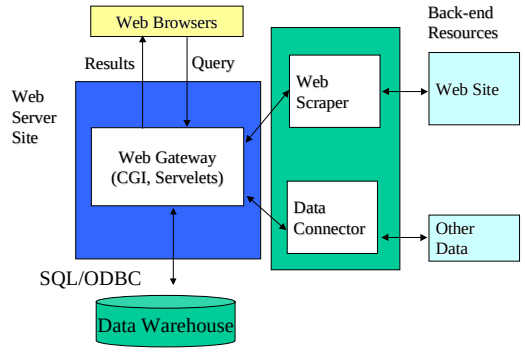


- "Data warehouse mediator" contains a global data dictionary
- Mediator can send the requests directly to the operational databases
- "Virtual" data warehouse (VDW) routes the user queries to the data sources (essentially a read-only distributed data management technology)
- VDW intelligence to automatically migrate data to the data warehouse (grow as you go)
- Plusses:
  - Flexibility, performance, scaling, and load balancing.
  - Can send local queries to regional and corporate queries corporate DW
  - Can support heterogeneity (one data mart may use a specialized DBMS suitable)
- Minusses:
  - Global data model may still be needed
  - For widely distributed data, performance degradation & service outages can occur.

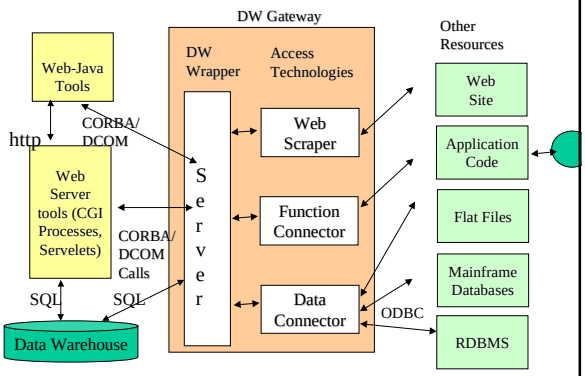
End User Design



Web Access to Data Warehouse



Integrated Access to DW from Web

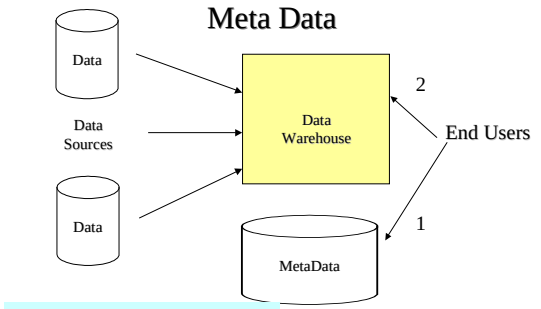


Meta Data

Meta Data, data about data, serves as a directory and a catalog of the data warehouse.

- Data warehouse must be accompanied with powerful and accurate meta data (more important than operational)
- Meta data can provide:
  - Data schema. Meta data shows the names, the attributes, the keys, and the formats of warehouse tables.
  - Semantic model. business objects and their relationships
  - Mapping of operational data to warehouse data.
  - Common routines for summarization and access of data
  - Predefined queries, reports, and spreadsheets.
  - Extract history.
  - Information about external and unstructured data.
  - Relationship to other meta data stores.
  - Data location (what tables are located where in the network).

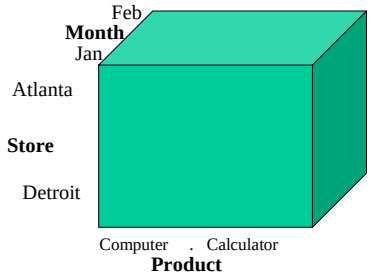
Meta Data



Steps in using the DW
1. Access metadata to learn about data
2. Issue the queries

Multidimensional Models and Databases

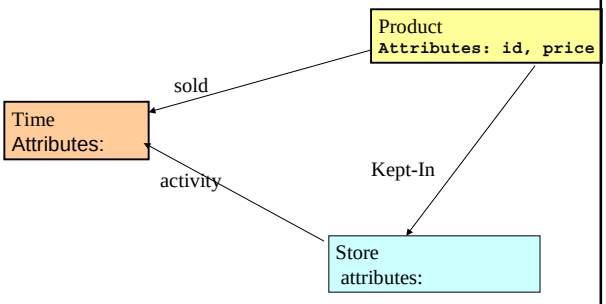
- Simple View: Multidimensional model is a three dimension spreadsheet (or more) (product, month, Store)



RDBMS is two dimensional

Main reference: "The data Warehouse Toolkit: Building Dimensional Data Warehouses", by Ralph Kimball, Wiley

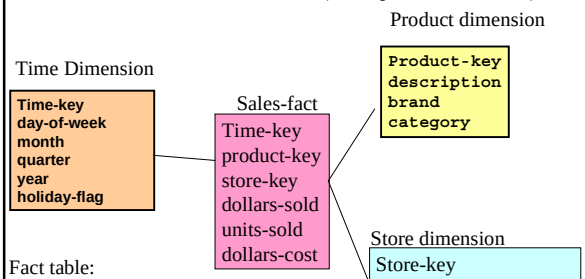
Entity Relationship Model



Issues with ER Models

- Each entity becomes a table
- Each relationship becomes a foreign key
- Can be very complex for queries
  - Some ER have hundreds of entitties (can cover wall of a room)
  - Normalization can generate more tables
  - Result: too many tables, interconnected in different ways to each other
- Suitable for OLTP (few tables accessed with no redundant data)
- Extremely difficult to conceptualize before issuing a query
- Not efficient for query processing (can result in large number of joins)

Dimensional Model (star join schema)



Fact table:
•daily item totals of all products sold
•one record per product per day (choice)
•can be very large
•can have several dimensions (inventory level, ship-to, customer, region,,)

Dimensional Model (Star-Join Schema)

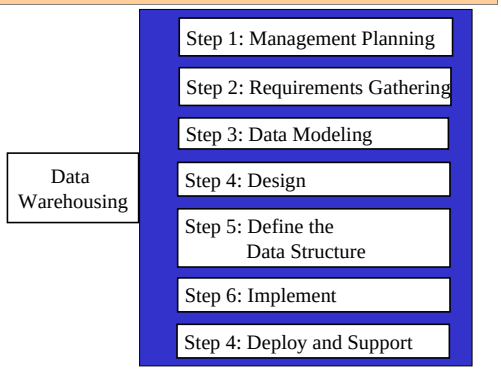
- Fact table:
  - Large dominant table in the middle
  - Only table with multiple joins to other tables
  - Can be designed at low or high granularity (for each product, daily activity, etc)
  - Can be very large (heavily indexed)
  - Stores numerical measurements of business
  - Facts should be numeric, continuously valued, and additive
  - Can be sparse (not every product is sold everyday)
- Dimension tables
  - Contain textual (descriptive) info
  - Attached to fact table through a single join
  - Should mostly consist of descriptions:
    - characters
    - numeric values (e.g., prices, sizes) that are used as descriptions
  - Attributes used for column headings

Multidimensional Databases (cont.)

- Multidimensional databases capture and present data as arrays that can be arranged in multiple dimensions,
- Multidimensional databases present large amounts of data to users in a manner that is easily comprehensible.
- Many EISs such as Pilot Software's Lightship and Comshare's Commander use mult-dimensional databases.
- These databases can be used to answer queries that would be extremely difficult cumbersome in SQL, such as

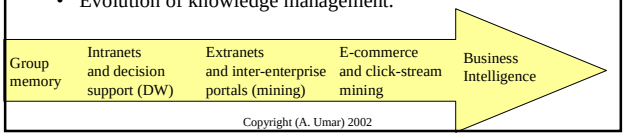
  "List the top five sales regions based on the percentage increase in revenues this year relative to last year".
- Use of RDBMS technology at the core of a multidimensional database (e.g., Microstrategy DSS Agent)
- Limitation of multidimensional databases: cannot store large data amounts (typically more than 10 GB of data)

Data Warehouse Development Process

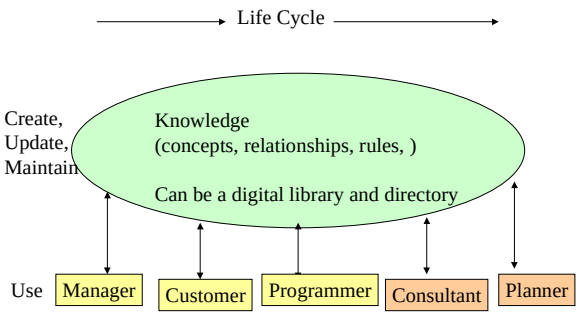


Business Intelligence and Knowledge Management

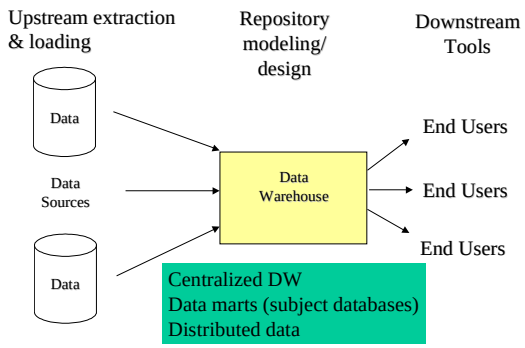
- Knowledge Management - capturing, managing, and using the corporate knowledge
- Business intelligence - Converting data into knowledge for business competition
- Levels: data, information, knowledge, wisdom
- Key idea: use in making/supporting decisions
- One person's data is someone else's knowledge
- Evolution of knowledge management:



Knowledge Management



Data Warehouse



Data Mining Overview

- **What is data mining**: Finding and discovering trends and patterns in data
- Also known as knowledge discovery KDD (knowledge discovery in databases)
- Emphasis is on hidden patterns and relationships
- Heavy use of AI and statistical analysis (more AI)
- Data to be mined is extracted and loaded in a mining file
- Examples:
  - What products the customer is likely to buy (based on current purchases)
  - Which customers are likely to discontinue services
  - What variables determine the customers who will go to a competitor
  - Who is most likely to respond to mail
  - How to detect fraudulent behaviour in credit card users
  - Which emarkets are likely to succeed

Issues with Data Mining

Plusses:

- Tremendous value to organizations (supporting businesss strategies)
- Data long forgotten is finding value ("hidden gold")
- Interesting applications of AI and statistics (many startups)
- Others

Minusses:

- Privacy (credit card, medical history, addresses, phone nos, people you call, Web sites you visit)
- Data ownership (who owns your data)
- Some attempts at regulations in US and Europe (e.g., cannot base models on gender, banks cannot give out credit history,etc.)
- Opinion: Too much reliance on past data, less original and innovative thinking
- Others?

Web Mining

Objective: Look for eCommerce events:

- Associated with a single user during a single visit.
- Look for Product events or Visit events.
- Personalization facilitates conversion of browsers to buyers.

Web Mining: Tools (Sample)

www.accrue.com www.webtrends.com www.netgen.com www.mineit.com

www.customerconversion.com www.angoss.com www.bluemartini.com

Components of a Web Site

A web-site is an interconnected network for presenting information where each web page is a node. Users navigate the nodes to discover information in ways that can be patterned and grouped. The assumption is that the identified patterns reveal useful and usable information about user preferences and motivations.

- Web-site content
- Web-site design
- Web-page content
- Web-page design

Web Mining: Features - Product Events

- View product (initial interest)
- Click-through (more information requested)
- Add to shopping cart
- Remove from shopping cart
- Buy / Bid / Order / Play (domain dependent)

Web Mining: Features - Visit Events

- Begin session or visit
- End session or visit
- Login
- Logout
- Personalization rule fire
- Service failure

Key Data Mining References

- Books
  - Groth, R., "Data Mining: Building Competitive Advantage", Prentice Hall, 2000
  - "Data Warehousing and Data Mining : Implementing Strategic Knowledge Management" by Elliot King, computer technology research
  - "Business Intelligence : The IBM Solution" by Mark Whitehorn, Mary Whitehorn. (1999)
  - "Building Data Mining Applications for CRM" by Alex Berson, et al. 1999
  - "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations" by Ian H. Witten, Eibe Fran
  - "Advances in Knowledge Discovery and Data Mining" by Usama M. Fayyad(Editor), et al.
- Annual conference, KDD (Knowledge Discovery and Data Mining): ACM SIGKDD
- White papers by many data mining vendors. Examples:
  - www.bluemartini.com
  - www.netgen.com

Key Points

- Data warehouses can be used for integration by keeping needed legacy data
- Many data warehouse architectures
- Web Data Warehousing is popular
- Meta data and Data Modeling
- Dimensional Data Modeling
- Data mining and web mining use data warehouses