UNIT-II

Data Warehousing Tools and Technology

Data Warehousing Tools

- **Any Data Warehousing Tool must have**
  - **Complexity of the data transformation**
  - **Strong data cleansing functionality**
  - **Data volume, means must have features that can speed up data movements**
  - **Therefore data warehousing tools must have large volume of data transferred.**

Characteristics

- The Tools must have the following characteristics

1. **Functional Capability**
2. **Ability to read directly from your data source**
3. **Metadata**

Cont…

1. **Functional Capability:**

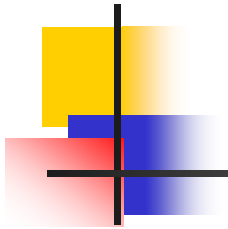
- This includes both the "**Transformation**" piece and the "**Cleansing**" piece.
- In general the typical ETL tools are either geared towards strong transformation capabilities or having strong cleansing capabilities
- but they are tied very strong in both, as a result ETL tool have strong cleansing capability.

Cont…

- **Ability to read directly from your data source:**
  - **For each organization , there is a different set of data sources**
  - **Make sure that ETL Tool should connect directly to your source data.**
- **Metadata:**
  - **The ETL tools plays a key role in your metadata**

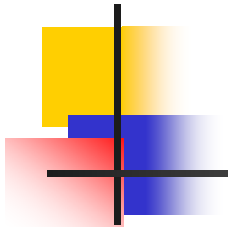
Cont..

- **Because it maps the source data to destination, which is an important piece of the metadata.**
- **Some organization have come to rely on the documentation of their ETL tools as their metadata source, as a result , it is very important to select an ETL tool that works with your overall metadata strategy.**



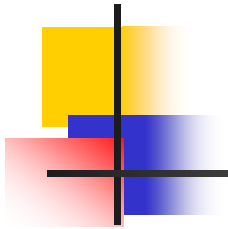
ETL Tool

- Extract, Transform and Load (ETL) is a process in data warehousing that involves.
- **Extracting** data from out sources.
- **Transforming** it according to business needs(which can include quality levels) and ultimately....
- **Loading** it in to the end target, that is the data warehouse.



Cont…

- These are sophisticated GUI based applications that enable the operation of **Extracting** data from source systems, **transforming** and **cleaning** data and **loading** in to the data warehouse.



Cont....

- ETL is important, as it is the way data actually gets loaded in to the warehouse.
- ETL is the process that loads any database
- ETL can also be used for the integration with legacy systems.
- Most popular tools are:
  - Informatica
  - Cognos
  - BI

ETL Tool

- **Extract: The first part of an ETL process is to extract the data from the source systems.**
- **Extraction converts the data in to a format for transformation processing**
- **An essential part of the extraction is the verification of extracted data, which should meets an expected pattern or structure.**
- **If not, the data is rejected.**

Cont…

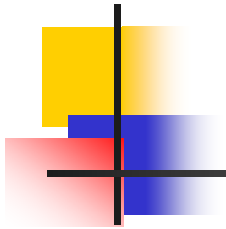
- **Transform:**
  - **The transform stage applies a series of rules or functions to the extracted data from the source to derive the data to be loaded to the end target.**
  - **Some data sources will require very little or even no manipulation of data.**
  - **In other cases one or more of the following transformation types to meet the business and technical needs of the end target may required.**

Cont…

- Translating coded values
- For eg. If the source system stores 1 for female and 2 for male , but warehouse stores M for male and F for female. This is called Automated Data Cleaning.
- That is encoding free-from values
- No manual cleaning occurs during ETL.
- Deriving new calculated value. (eg. Sale_amount = qty*unit_price)

Cont…

- **Joining together data from multiple sources (eg. Lookup, merge etc.)**
- **Summarizing multiple rows of data. (eg. Total sales for each row and for each region)**
- **Transposing or pivoting (turning multiple columns in to multiple rows or vice versa)**
- **Splitting a column in to multiple columns.**
- **Applying any form of simple or complex data validation**



Load

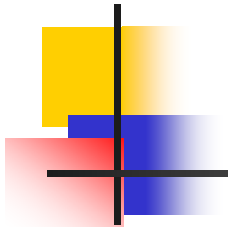
- The **load phase loads the data in to the end target that is Data warehouse**.
- Depending on the requirements of the organization, this process ranges widely.
- Some data warehouse might be overwrite existing information with cumulative , updated data
- While other DW might add new data in a historical form

Cont...

- The timing and the scope to replace and upend are strategic design choices dependent on the time available and the business needs.
- More complex systems can maintain a history and audit trail of all changed to the loaded in the warehouse.
- As the **load phase** interact with a database, the constraints in the database schema as well as in triggers activated upon data load apply.
- It also contributes to the overall data quality performance of the ETL process.

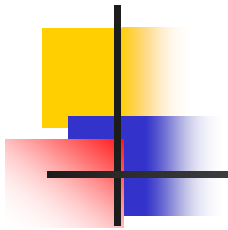
Data warehouse database

- The **data warehouse** and the **OLTP** data base are both **relational databases**. However, the objectives of both these databases are different.
- Data warehouse databases are designed for **analysis of business** measures by categories and attributes
- Optimized for bulk loads and large, complex, unpredictable queries that access many rows per table.
- Loaded with consistent, valid data; requires no real time validation



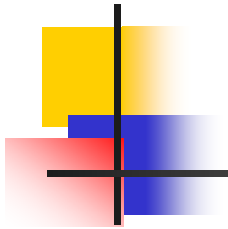
Cont…

- Supports **few concurrent users** relative to OLTP
- The **data warehouse** does not supply to real time operational requirements of the enterprise.
- **It is more a storehouse of current and historical data and may also contain data extracted from external data sources.**



OLTP Database

- It is also **relational database**.
- The **OLTP database** records transactions in real time and aims to automate clerical data entry processes of a business entity. Addition, modification and deletion of data in the OLTP database is essential and the semantics of the application used in the front end impact on the organization of the data in the database.

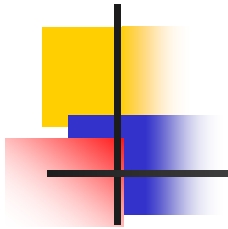


OLTP database

- Designed for real time business operations.
- Optimized for a common set of transactions, usually adding or retrieving a single row at a time per table.
- Optimized for validation of incoming data during transactions; uses validation data tables.
- Supports thousands of concurrent users.

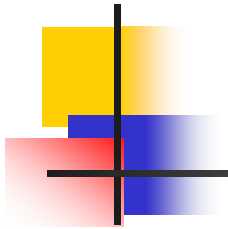
Data Warehouse Metadata

- Data warehouse metadata are **pieces of information** stored in one or more special-purpose metadata repositories.
- That include information on the contents of the data warehouse.
- Their location and their structure, information on the processes that take place in the data warehouse back-stage.



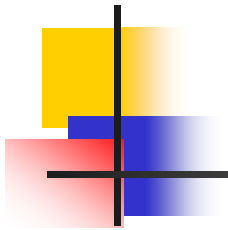
Cont...

- concerning the refreshment of the warehouse with clean, up-to-date, semantically and structurally reconciled data.
- Information on the implicit semantics of data (with respect to a common enterprise model), along with any other kind of data that aids the end-user exploit the information of the warehouse.



Cont...

- Information on the infrastructure and physical characteristics of components and the sources of the data warehouse. &
- Information including security, authentication, and usage statistics that aids the administrator tune the operation of the data warehouse as appropriate.

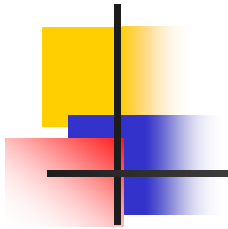


Cont…

- Data warehouse metadata repositories store large parts of this kind of *data warehouse metadata.*

Data warehouse Management Tools

- Data Warehouse **Management Tools** are **software applications** that extract and transform data from operational systems and loads it into the data warehouse**.**
- The area of data warehouse management is very complex.
- In this data captured from operational data sources such as those data coming from transactional business software solutions like Supply Chain Management (SCM)



Cont...

Point of Sale, Customer Serving Software and Enterprise Resource Planning (ERP) and management software to undergo the ETL (extract, transform, load) process.

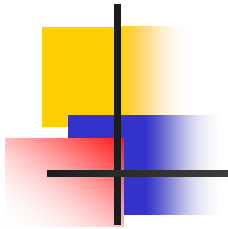
UNIT-III

Data Warehouse Design

Data Warehouse Design

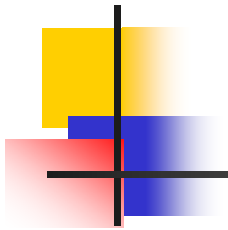
- For designing a data warehouse
- Starting with the **need** and **motivation,** the sequential steps involved for developing a data warehouse are.....

1. **Content in a data warehouse**
2. **Hardware**
3. **Networking infrastructure requirement for the data warehouse**
4. **The soft environment and various Tools**



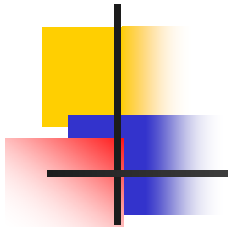
Cont...

5. The issues involved in **multiprocessor architecture** and **I/O devices** have been discussed.
6. The **range of tools available** in the market with their **mutual compatibility** requirements and their features and capabilities.



Design of Data Warehouse

- The job of designing and implementing a data warehouse is very challenging and difficult.
- There are so many questions while designing a data warehouse.
- Where to start?
- Which data should put first?
- Where is the data available?
- Which query should be answered?



Cont...

- How would you bring down the scope of the project to something smaller and manageable?
- How it would be scalable to gradually upgrade to build a comprehensive data warehouse environment?

Cont...

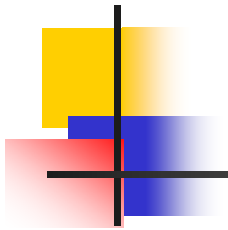
- A data warehouse can be built either on a **top-down or a bottom-up approach.**
- We can design a global data warehouse for an entire organization and split it up in to individual **data marts** or **sub data warehouses** dedicated for individual departments.
- Alternatively individuals **data marts** can be built for each department and finally they all get integrated in to a **central data warehouse.**

Top – Down Approach

- In top – down approach we build up first **Meta Data** then **Data Mart** and finally a **Data Warehouse .**
- **The recent trend is to build data mart before a large data warehouse is built.**
- People want something smaller, so as to get manageable**.**

Bottom-up Approach

- In the **bottom-up approach** is more realistic.
- Here we design and create the **Data Warehouse** first then **data marts** and finally **Metadata**.
- but the integration of individual data mart should be made easier with advanced planning and preparation.
- In building a data warehouse, the organization has to make arrangements for information flow from various Internal Information Systems and databases as well as from internal sources of information.

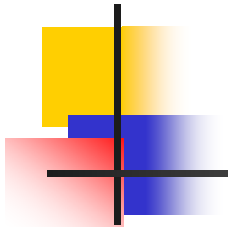


Cont...

- This requires close involvement of the users in identifying the information requirements and identifying sources of the same from both internal and external data sources and information system in time.

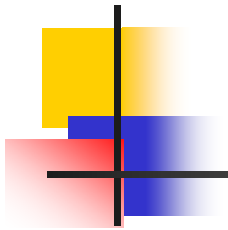
Important steps to design a Data Warehouse

1. Choose a subject matter.(one at a time)
2. Decide what the fact table represents
3. Identify and confirm the dimensions
4. Choose the facts
5. Store pre-calculations in the fact table
6. Define the dimensions and table
7. Decide the duration of the database and periodicity of updation
8. Track slowly the changing dimensions
9. Decide the query properties and the query methods



Cont…

- All the above steps are required before the data warehouse is implemented.
- The final step 10 is to implement a simple data warehouse or data mart.
- **First only the data marts are identified, designed and implemented.**
- **Secondly the data warehouse will come out gradually.**



Implementation

- A data warehouse can not be purchased and installed.
- Its implementation requires the integration of implementation of many products.

Followings are the steps of the Data Warehouse Implementation

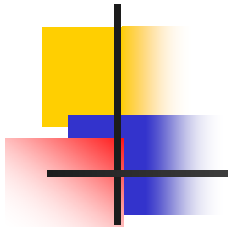
Step 1. Collect and analyze business requirements

Step 2. Create a data model and physical design for the data warehouse after deciding the appropriate hardware platform.

Step 3. Define the data sources.

Step 4. Choose the DBMS and software platform for data warehouse.

Step 5. Extract the data from operational data sources, transform it , clean-up and load in to the data warehouse model or data mart.



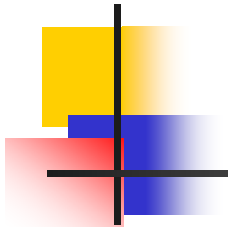
Cont...

- **Step 6.** Choose database access and reporting tools.
- **Step 7.** Choose database connectivity software.
- **Step 8.** Choose data analysis (OLAP) and presentation software (client GUI).
- **Step 9.** Keep refreshing the data warehouse periodically

Major Issues in Data Warehouse

- There are **three major issues** that will be faced in data warehouse development.

1. Heterogeneity of data source requiring substantial efforts in data conversion.(data source to data conversion)
2. Maintaining timeless and high-quality levels of data integrity, reliability and authenticity. Further data may be quit old and historical, while old data of past is essential for a data warehouse, but it will be relevant and useful in the data warehouse form.

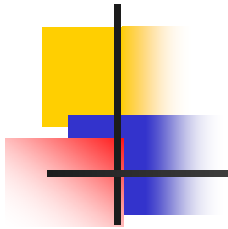


Cont...

3. Another issue is the tendency of the data warehouse to grow very large.

   So discrete decisions should be made by the designer of the data warehouse in limiting the size of the warehouse

PROJECT PLANNING AND MANAGEMENT



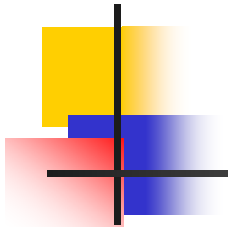
OBJECTIVES

- Review the essentials of planning for a data warehouse
- Distinguish between data warehouse projects and OLTP system projects
- Learn how to adapt the life cycle approach for a data warehouse project
- Discuss project team organization, roles, and responsibilities
- Consider the warning signs and success factors

Introduction

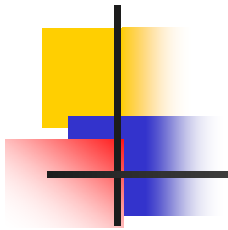
- **This chapter is designed not just for the project manager or the project coordinator, it is designed for all IT professionals irrespective of their roles in data warehousing projects.**
- **This chapter will show you how best you can fit into your specific role in a project. If you want to be part of a team that is passionate about building a successful data warehouse, you need.**

TRENDS IN DATA WAREHOUSING



Objectives

- Continuous growth in data warehousing
- Learn how data warehousing is becoming mainstream
- Major trends
- Need of standards and review the progress
- Web-enabled data warehouse



Introduction

- In DW trends we will see
- **How and why data warehousing continues to grow** and become more and more omnipresent.
- We will also discuss the trends in vendor solutions and products.

CONTINUED GROWTH IN DATA WAREHOUSING

- Data warehousing is no longer a purely new idea for study and experimentation.
- It is becoming mainstream.
- Every company that has a DW is realizing enormous benefits giving positive results.
- Many companies now incorporating **Web-based technologies,** are enhancing the potential for greater and easier delivery of vital information.
- Over the past five years, hundreds of vendors have flooded the market with numerous products. Vendor solutions and products run the scope of data warehousing:
- **data modeling**
- **data quality**
- **data analysis**
- **metadata and so on**

Data Warehousing is Becoming Mainstream:

Four Significant Factors droves many companies to move in to Data Warehousing:

- **severe competition**
- **Government deregulation**
- **Need to accomplish internal processes**
- **necessary for customized marketing**

Data Warehouse Expansion(keeping summary data for high-level analysis)

Vendor Solutions and Products

first ones to adopt data Warehousing

- Telecommunications
- Banking and
- Retail

That was largely because of government deregulation in **telecommunications and banking.**

Retail business moved in to data Warehousing because of severe competition.

Then the business get in to the DW consisted of companies in **financial services, health care, insurance, manufacturing, pharmaceuticals, transportation and distribution.**

Vendor Solutions and Products

There are hundreds of DW vendors and thousands of DW products and solutions. Fig. 3.3 shows a list of products, grouped by functions they performed in DW.

PRODUCTS BY FUNCTIONS (Number of leading products shown within parenthesis

- Data Integrity & Cleansing (12)
- Data Modeling (10)
- Extraction/Transformation
  - *Generic (26)*
  - *Application-specific (9)*
- Data Movement (12)
- Information Servers
  - *Relational DBs (9)*
  - *Specialized Indexed DBs (5)*
  - *Multidimensional DBs (16)*
- Decision Support
  - *Relational OLAP (9)*
  - *Desktop OLAP (9)*
  - *Query & Reporting (19)*
  - *Data Mining (23)*
  - *Application Development (9)*
- Administration & Management
  - *Metadata Management (14)*
  - *Monitoring (5)*
  - *Job Scheduling (2)*
  - *Query Governing (3)*
  - *Systems Management (1)*
- DW Enabled Applications
  - *Finance (10)*
  - *Sales/Marketing/CRM (23)*
  - *Balanced Scorecard (5)*
  - *Industry specific (21)*
- Turnkey Systems (14)

Source: The Data Warehousing Institute

Figure 3-3 Data warehousing products by functions.

SIGNIFICANT TRENDS

This is the beginning to see important progress in the software specially for optimizing queries, indexing very large tables, enhancing SQL, improving data compression and Expanding dimensional modeling:

- Multiple Data Types
- Searching Unstructured Data.
- Spatial Data.(ex. address, street block, city, country, state and zone)
- Data Visualization
- Parallel Processing
- Query Tools
- Browser Tools (allowing users to brows the data dictionary or metadata, finding an informational object of interest)
- Data Fusion (data from numerous sources are integrated)
- Multidimensional Analysis
- Agent Technology (software agents is a program which can be used to sort and filter out e-mail according to rules defined by the users)
- Active Data Warehousing (opening your DW to 30,000 users worldwide, consisting of employees, customers and business partners, in addition to allowing

Decision Making and Data Warehousing

"A data warehouse is the data, processes, tools, and facilities to manage and deliver complete, timely, accurate, and understandable business information to authorized individuals for effective decision making."

- **Structured Data**
  - Includes traditional relational databases
  - Typically internal and enterprise-owned
  - Predetermined
- **Unstructured Data**
  - Includes articles, reports, images, and videos
  - Utilizes external data and expert opinion
  - Ad hoc

Multiple Data Types

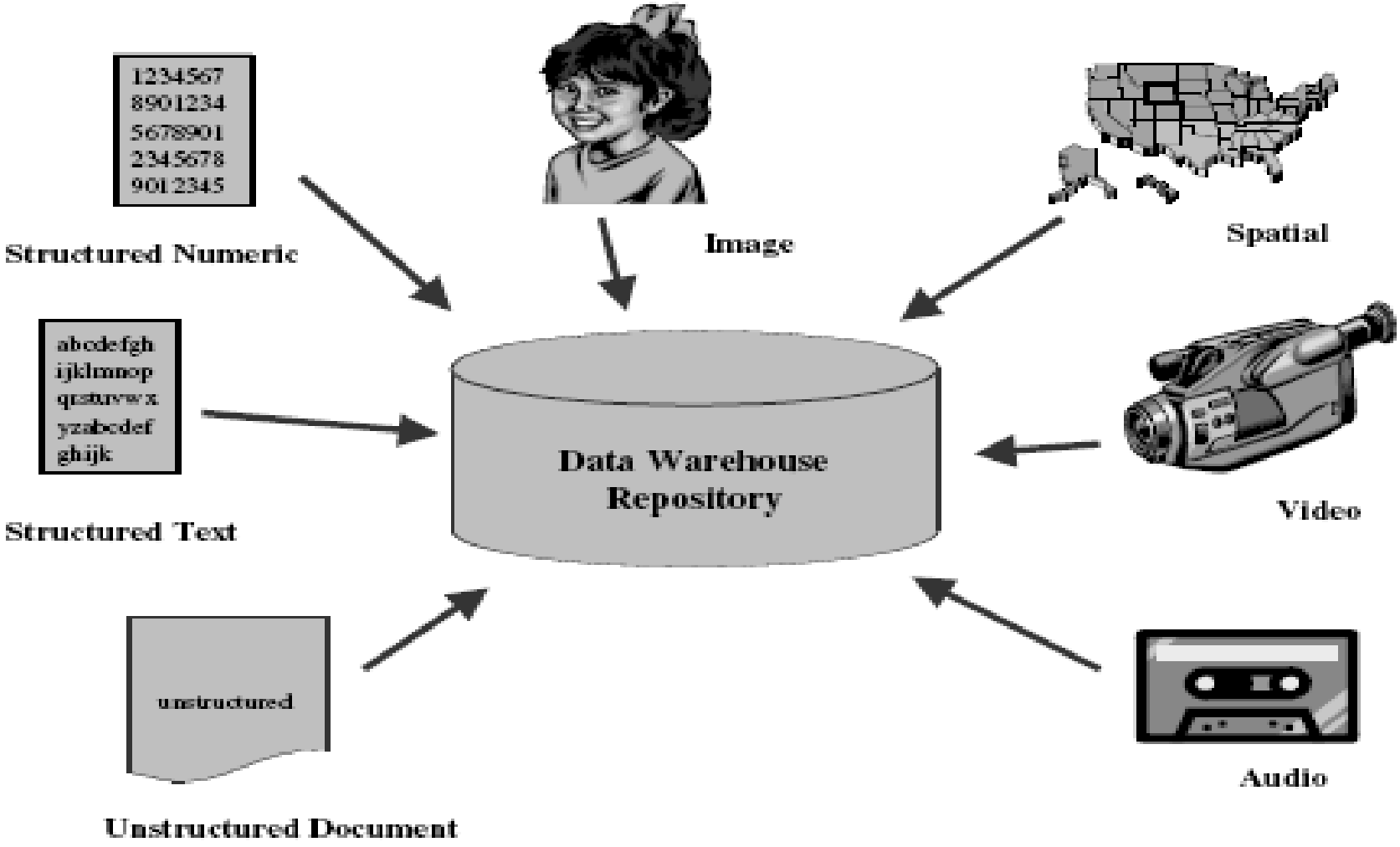


Figure 3-4 Data warehouse: multiple data types.

EMERGENCE OF STANDARDS

- In each of the multitude of technologies supporting the data warehouse, numerous vendors and products exist. The implication is that when you build your data warehouse, many choices are available to you to create an effective solution with the best-of-breed products. That is the good news. However, the bad news is that when you try to use multivendor products, the result could also be total confusion.
- **Metadata :** Two separate bodies are working on the standards for metadata: **the Meta Data Coalition and the Object Management Group**
- ***Meta Data Coalition.: Working on a standard Open Information Model (OIM) and joined with Microsoft ++ in December 1998***
- ***The Object Management Group. :*** *Oracle, IBM,* Hewlett-Packard, Sun, and Unisys sought
- **OLAP : The OLAP Council was established in** January 1995.

WEB-ENABLED DATA WAREHOUSE

- They are of two types:

1. The Warehouse to the Web
2. The Web to the Warehouse (Web house)

- On one hand we can transform DW in to Web-enabled DW or We have to bring DW to the web
- On other hand we need to bring the web to DW

The Web house can produce the following useful information:

- Site statistics
- Visitor conversions
- Ad metrics
- Referring partner links
- Site navigation resulting in orders
- Site navigation not resulting in orders
- Pages that are session killers
- Relationships between customer profiles and page activities
- Best customer and worst customer analysis

SUMMARY

- Data warehousing is becoming mainstream with the spread of high-volume data warehouses and the rapid increase in the number of vendor products.
- To be effective, **modern data warehouses need to store multiple types of data: structured and unstructured, including documents, mages, audio, and video.**
- **Data visualization deals with displaying information in several types of visual forms: text, numerical arrays, spreadsheets, charts, graphs, and so on. Tremendous progress has been made in data visualization.**
- **Data warehouse performance may be improved by using parallel processing with appropriate hardware and software options.**
- It is critical to adapt data warehousing to work with ERP packages, knowledge management, and customer relationship systems.
- **Data warehousing industry is seriously seeking agreed-upon standards for metadata and OLAP.**
- **Web-enabling the data warehouse means using the Web for information delivery and integrating the click stream data from the corporate Web site for analysis**.
- The convergence of data warehousing and the Web technology is crucial to every business in the 21st century.

Trends In Data Warehousing

Data Warehousing is Becoming Mainstream

In the early stages, four significant factors drove many companies to move into data warehousing:

- violent competition
- Government deregulation
- Need to restore internal processes
- Imperative for customized marketing

Significant Factors

These significant factors reflect the new trends in data warehousing:

- Multiple Data Types
- Data Visualization
- Parallel Processing
- Query Tools
- Browser Tools
- Data Fusion
- Multidimensional Analysis
- Agent Technology
- E-Business- ERP, KM, CRM

Decision Making and Data Warehousing

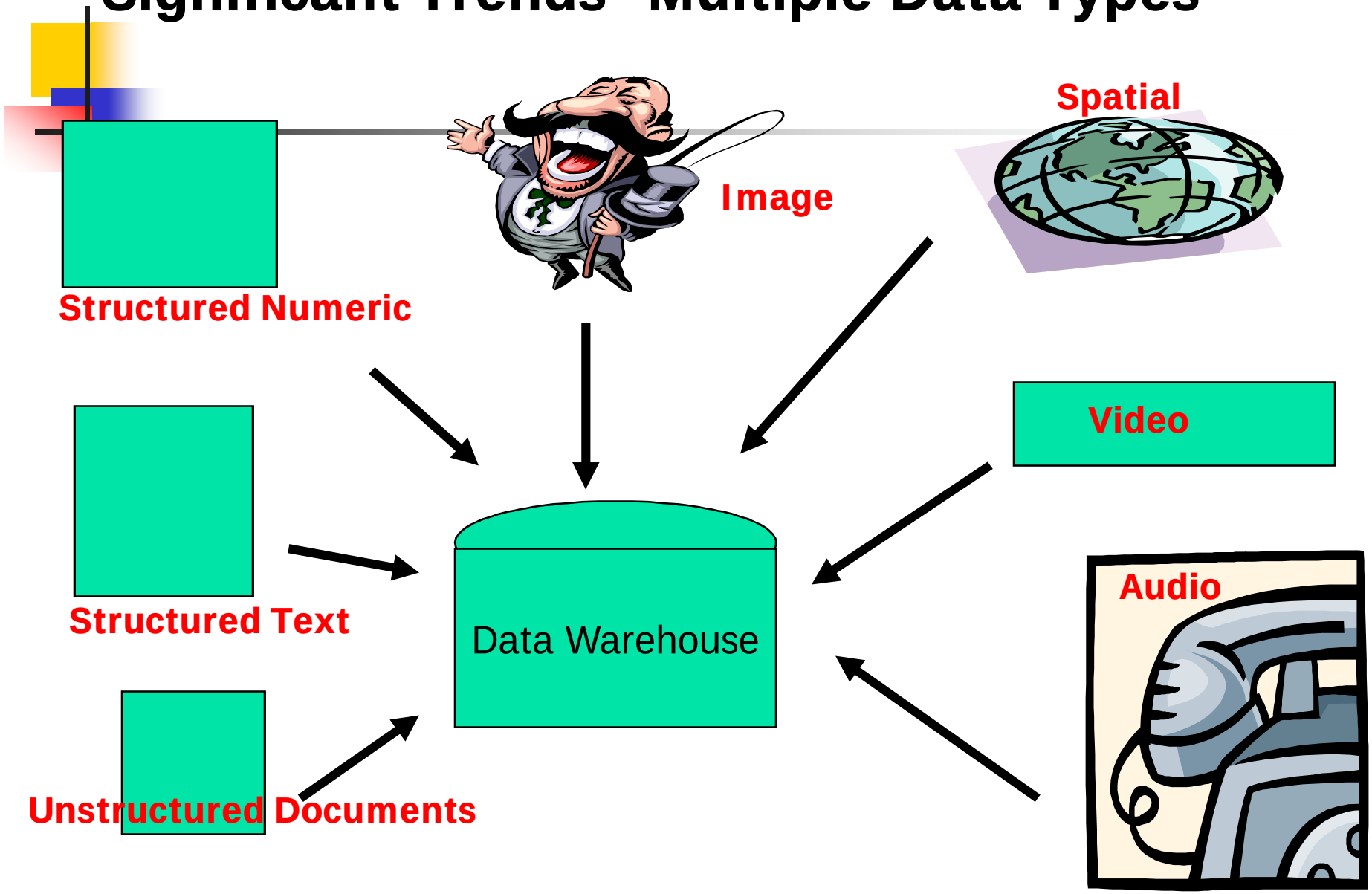
"A data warehouse is the data, processes, tools, and facilities to manage and deliver complete, timely, accurate, and understandable business information to authorized individuals for effective decision making."

- **Structured Data**
  - Includes traditional relational databases
  - Typically internal and enterprise-owned
  - Predetermined
- **Unstructured Data**
  - Includes articles, reports, images, and videos
  - Utilizes external data and expert opinion
  - Ad hoc

Decision Making and Data Warehousing

- **Management Systems**
  - Extend relational databases to store and support multimedia
  - User-defined types (UDT) and functions (UDF) in SQL-3
- **Specialized Servers**
  - Used for data which is incompatible with relational databases (e.g., Streaming video servers)
  - Objects may be linked to a relational database
- **Search Engines**
  - Query by Image Content (shape, color, texture, etc)
  - Text retrieval on free-text documents
  - Audio and video searching

Significant Trends- Multiple Data Types



17

Important Questions on DW

- What is Data Warehouse? Explain in detail.
- Draw and explain the three tier or Multitier Data Warehouse Architecture.
- Explain Data warehouse component with suitable diagram.
- What is OLAP? Explain OLAP operations along with its types.
- Explain Star Schema and snowflake Schema or Physical design process of Data Warehouse?
- What is multidimensional data model? Explain with neat diagram.

IMP QUE Cont .....

- Compare the OLTP and OLAP.
- What is Meta Data? Explain in detail.
- What is Data Marts? Explain in detail.
- Write and explain the complete designing steps of Data Warehouse.
- How to implement a Data Warehouse? Write and explain the complete steps of Implementation.
- What are the various trends in the data warehouse.
- What do you mean by project planning and requirement? Explain how it is necessary in DW.