Data Warehouse

1990

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

Subject Oriented:

Integrated:

Time-variant:

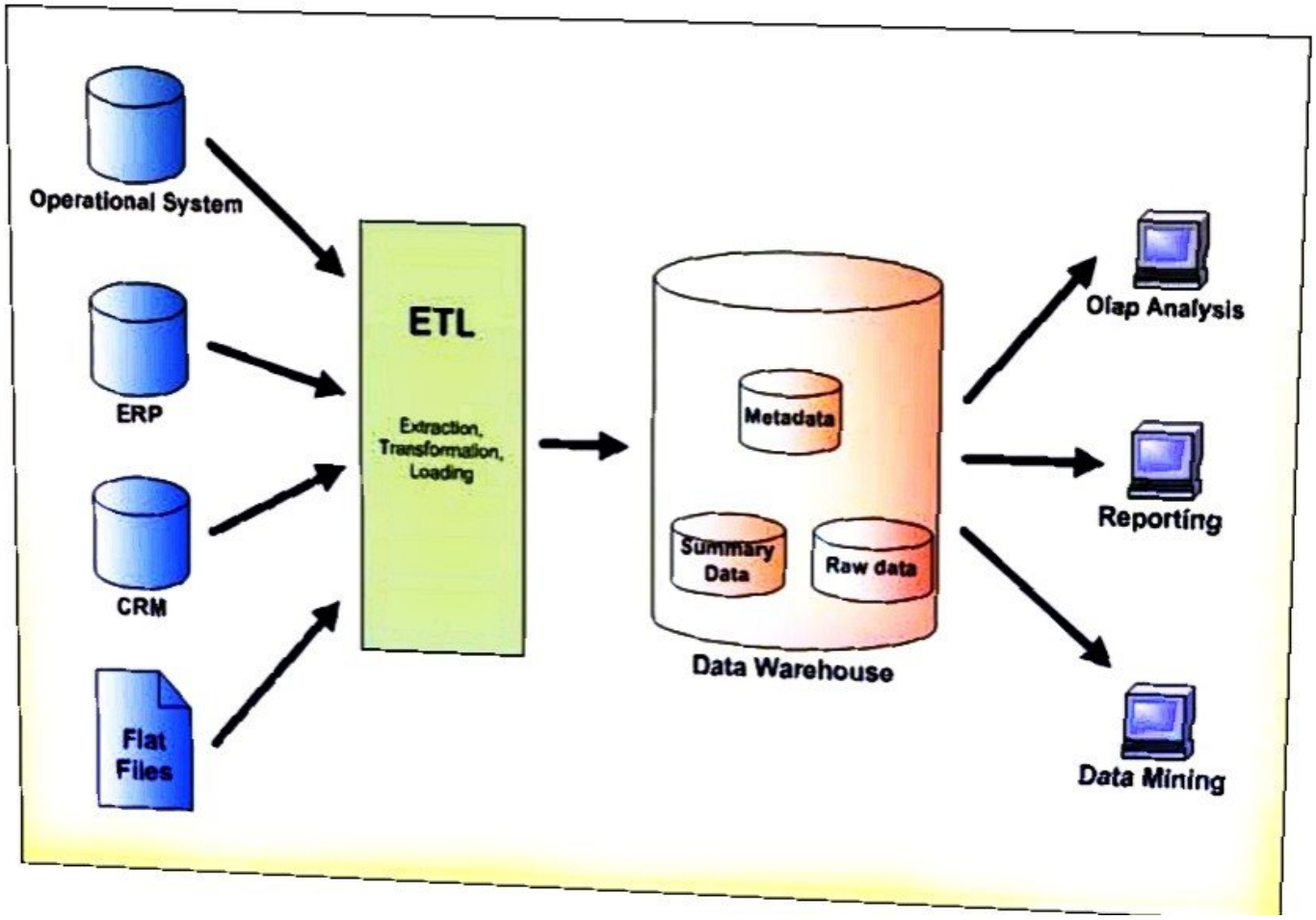
Non-volatile:

A data warehouse is a copy of transaction data specifically structured for query and analysis.

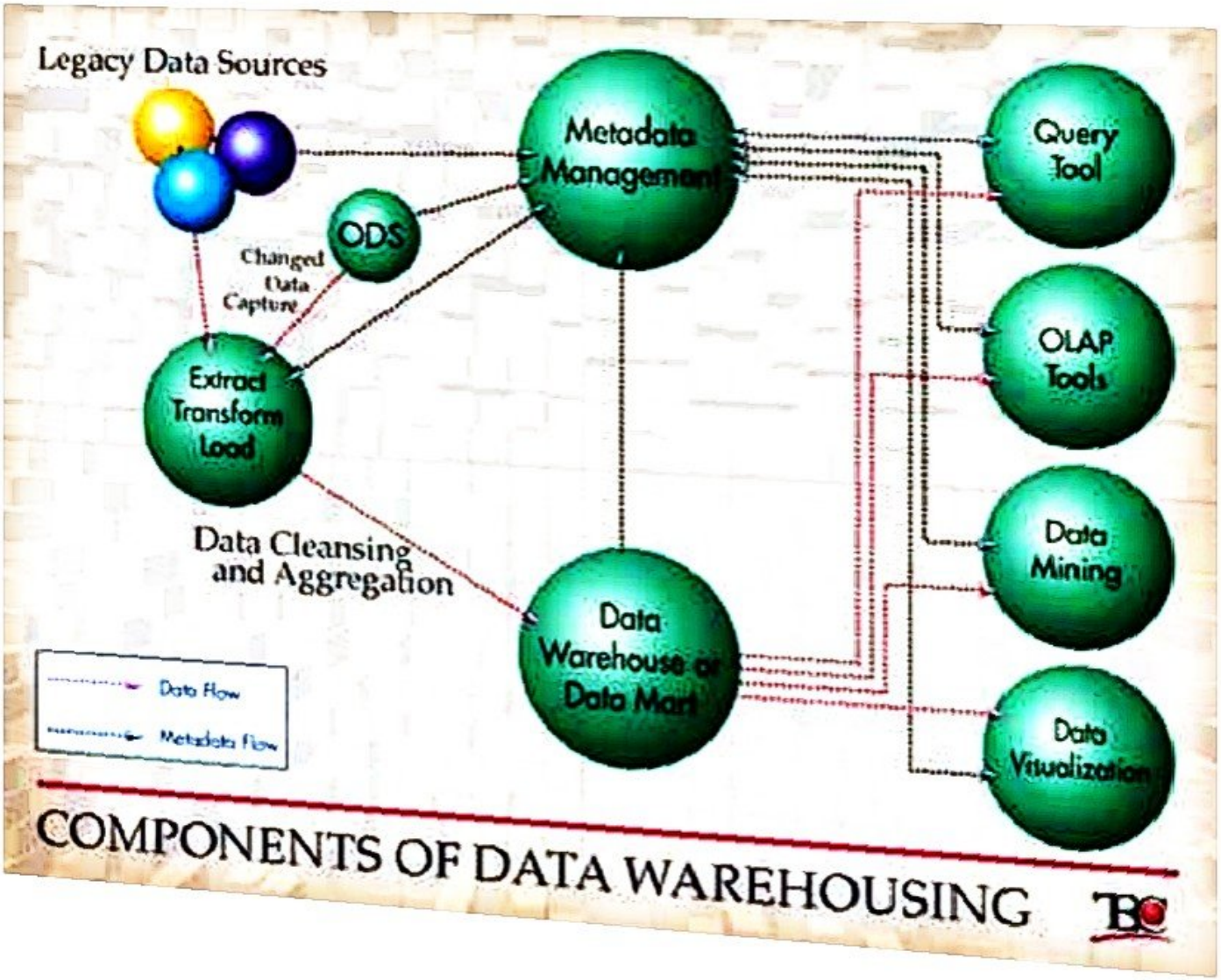
Another definition:

A data warehouse is a repository (data & metadata) that contains integrated, cleansed, and reconciled data from disparate sources for decision support applications, with an emphasis on online analytical processing. Typically the data is multidimensional, historical, non volatile.

Data Warehouse Architecture



Components of Data Warehousing

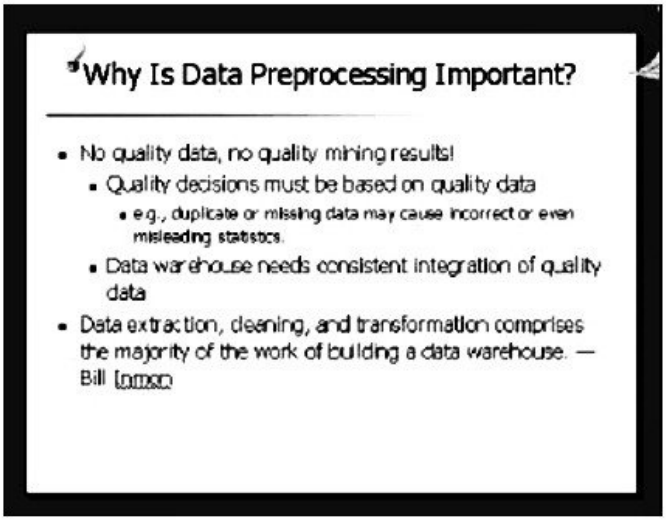


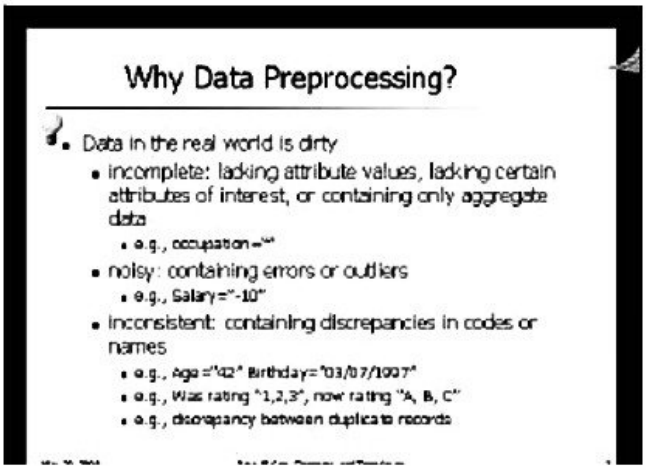
Discussions

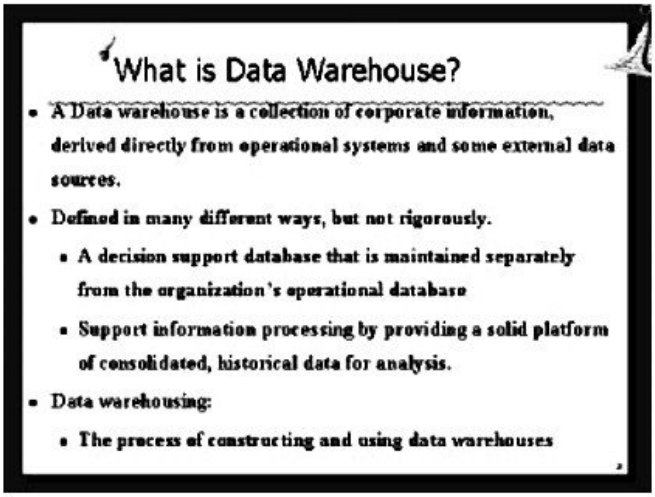
- Write short notes on:
  - Legacy systems
  - Data warehouse
  - Standard Business Applications
- What is a Data warehouse? How does it differ from a database?
- Discuss various factors, which lead to Data Warehousing.
- Briefly discuss the history behind Data warehouse.

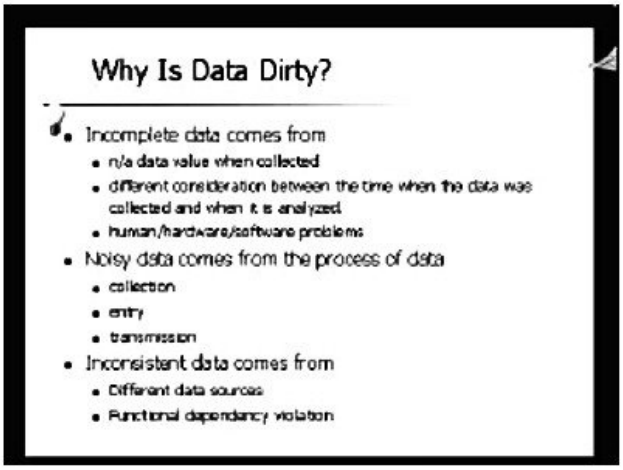
References

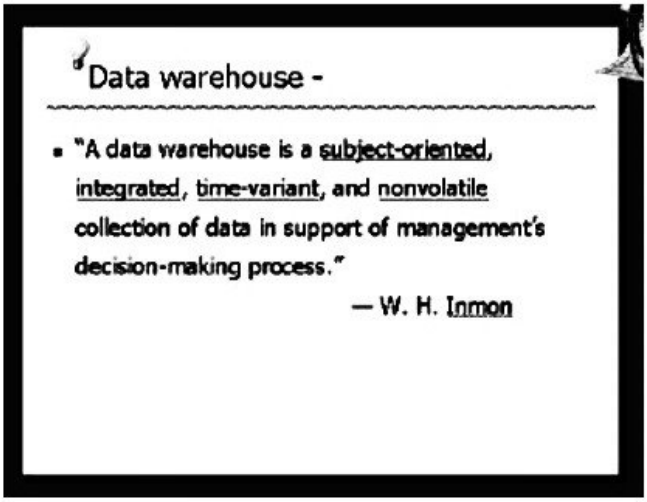
1. **Adriaans, Pieter,** *Data mining,* Delhi: Pearson Education Asia, 1996.
2. Anahory, Sam, Data warehousing in the real world: a practical guide for building decision support systems, Delhi: Pearson Education Asia, 1997.
3. Berry, Michael J.A. ; Linoff, Gordon, Mastering data mining : the art and science of customer relationship management, New York : John Wiley & Sons, 2000
4. Corey, Michael, Oracle8 data warehousing, New Delhi: Tata McGraw- Hill Publishing, 1998.
5. Elmasri, Ramez, Fundamentals of database systems, 3rd ed. Delhi: Pearson Education Asia, 2000.







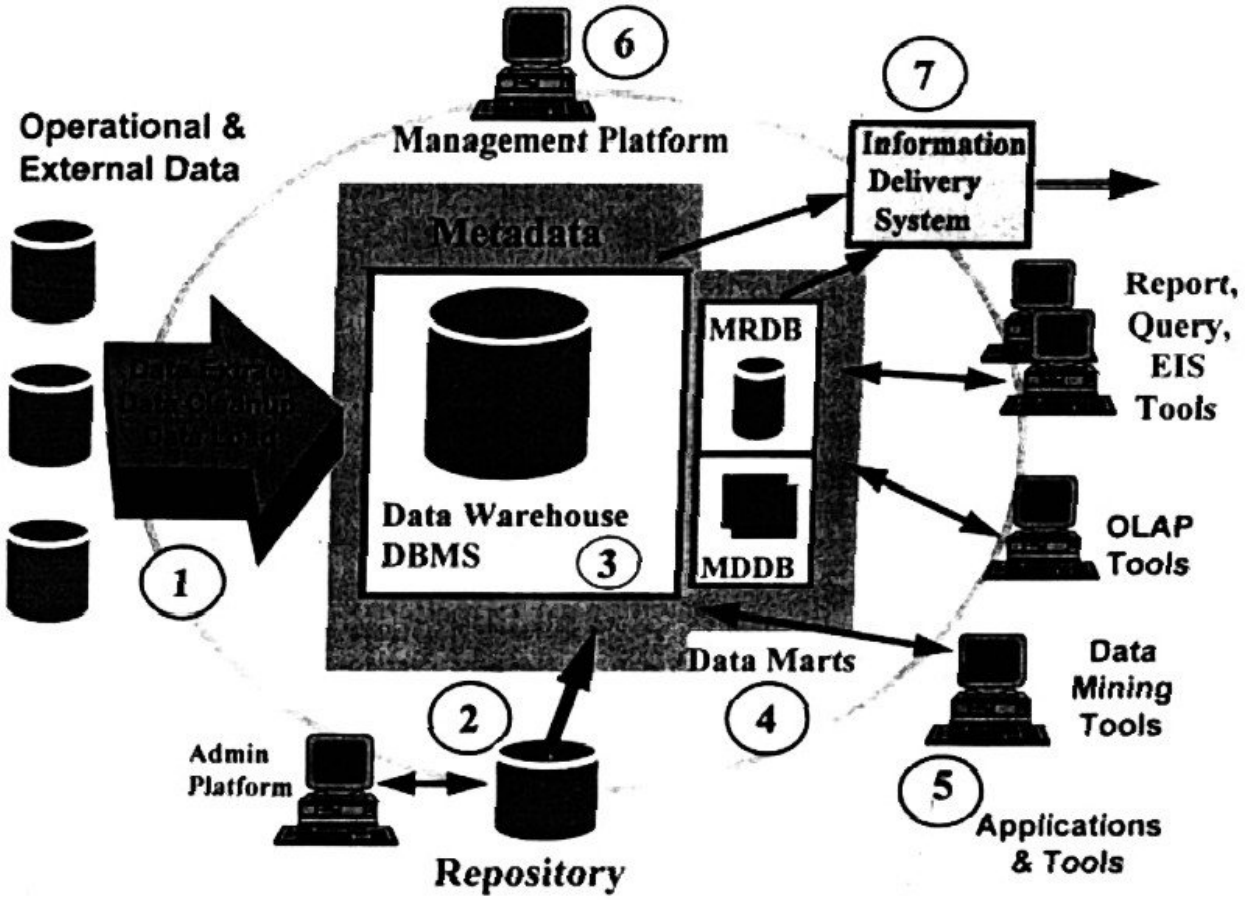




Data warehouse Architecture and its seven components

1. Data sourcing, cleanup, transformation, and migration tools
2. Metadata repository
3. Warehouse/database technology
4. Data marts
5. Data query, reporting, analysis, and mining tools
6. Data warehouse administration and management
7. Information delivery system

Data warehouse is an environment, not a product which is based on relational database management system that functions as the central repository for informational data. The central repository information is surrounded by number of key components designed to make the environment is functional, manageable and accessible.



Data warehouse environment.

The data source for data warehouse is coming from operational applications. The data entered into the data warehouse transformed into an integrated structure and format. The transformation process involves conversion, summarization, filtering and condensation. The data warehouse must be capable of holding and managing large volumes of data as well as different structure of data structures over the time.

1. Data warehouse database

This is the central part of the data warehousing environment. This is the item number 2 in the above arch. diagram. This is implemented based on RDBMS technology.

2. Sourcing, Acquisition, Clean up, and Transformation Tools

This is item number 1 in the above arch diagram. They perform conversions, summarization, key changes, structural changes and condensation. The data transformation is required so that the information can by used by decision support tools. The transformation

- Subject areas, and info object type including queries, reports, images, video, audio clips etc.
- Internet home pages
- Info related to info delivery system
- Data warehouse operational info such as ownerships, audit trails etc.,

Meta data helps the users to understand content and find the data. Meta data are stored in a separate data stores which is known as informational directory or Meta data repository which helps to integrate, maintain and view the contents of the data warehouse.

The following lists the characteristics of info directory/ Meta data:

- It is the gateway to the data warehouse environment
- It supports easy distribution and replication of content for high performance and availability
- It should be searchable by business oriented key words
- It should act as a launch platform for end user to access data and analysis tools
- It should support the sharing of info
- It should support scheduling options for request
- IT should support and provide interface to other applications
- It should support end user monitoring of the status of the data warehouse environment

4. Access tools

Its purpose is to provide info to business users for decision making. There are five categories:

- Data query and reporting tools
- Application development tools
- Executive info system tools (EIS)
- OLAP tools
- Data mining tools

Query and reporting tools are used to generate query and report. There are two types of reporting tools. They are:

Managed Query tools: used to generate SQL query. It uses Meta layer software in between users and databases which offers a point-and-click creation of SQL statement. This tool is a preferred choice of users to perform segment identification, demographic analysis, territory management and preparation of customer mailing lists etc.

Application development tools: This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all db systems

OLAP Tools: are used to analyze the data in multi dimensional and complex views. To enable multidimensional properties it uses MDDB and MRDB where MDDB refers multi dimensional data base and MRDB refers multi relational data bases.

Data mining tools: are used to discover knowledge from the data warehouse data also can be used for data visualization and data correction purposes.

5. Data marts

Departmental subsets that focus on selected subjects. They are independent used by dedicated user group. They are used for rapid delivery of enhanced decision support functionality to end users. Data mart is used in the following situation:

- Extremely urgent user requirement
- The absence of a budget for a full scale data warehouse strategy
- The decentralization of business needs
- The attraction of easy to use tools and mind sized project

Data mart presents two problems:

1. Scalability: A small data mart can grow quickly in multi dimensions. So that while designing it, the organization has to pay more attention on system scalability, consistency and manageability issues
2. Data integration

6. Data warehouse admin and management

The management of data warehouse includes,

Other important terminology

- *Enterprise Data warehouse:* It collects all information about subjects (*customers, products, sales, assets, personnel*) that span the entire organization
- Data Mart: Departmental subsets that focus on selected subjects. A data mart is a segment of a data warehouse that can provide data for reporting and analysis on a section, unit, department or operation in the company, e.g. sales, payroll, production. Data marts are sometimes complete individual data warehouses which are usually smaller than the corporate data warehouse.
- *Decision Support System (DSS):* Information technology to help the knowledge worker (executive, manager, and analyst) makes faster & better decisions
- *Drill-down:* Traversing the summarization levels from highly summarized data to the underlying current or old detail
- *Metadata:* Data about data. Containing location and description of warehouse system components: names, definition, structure…

Benefits of data warehousing

- Data warehouses are designed to perform well with aggregate queries running on large amounts of data.
- The structure of data warehouses is easier for end users to navigate, understand and query against unlike the relational databases primarily designed to handle lots of transactions.
- Data warehouses enable queries that cut across different segments of a company's operation. E.g. production data could be compared against inventory data even if they were originally stored in different databases with different structures.

Top - Down Approach

In the top down approach suggested by Bill Inmon, we build a centralized repository to house corporate wide business data. This repository is called Enterprise Data Warehouse (EDW). The data in the EDW is stored in a normalized form in order to avoid redundancy. The central repository for corporate wide data helps us maintain one version of truth of the data. The data in the EDW is stored at the most detail level. The reason to build the EDW on the most detail level is to leverage

1. Flexibility to be used by multiple departments.
2. Flexibility to cater for future requirements.

The disadvantages of storing data at the detail level are

1. The complexity of design increases with increasing level of detail.
2. It takes large amount of space to store data at detail level, hence increased cost.

Once the EDW is implemented we start building subject area specific data marts which contain data in a de normalized form also called star schema. The data in the marts are usually summarized based on the end users analytical requirements. The reason to de normalize the data in the mart is to provide faster access to the data for the end users analytics. If we were to have queried a normalized schema for the same analytics, we would end up in a complex multiple level joins that would be much slower as compared to the one on the de normalized schema.

We should implement the top-down approach when

1. The business has complete clarity on all or multiple subject areas data warehouse requirements.
2. The business is ready to invest considerable time and money.

The advantage of using the Top Down approach is that we build a centralized repository to cater for one version of truth for business data. This is very important for the data to be reliable, consistent across subject areas and for reconciliation in case of data related contention between subject areas.

The disadvantage of using the Top Down approach is that it requires more time and initial investment. The business has to wait for the EDW to be implemented followed by building the data marts before which they can access their reports.

Bottom Up Approach

The bottom up approach suggested by Ralph Kimball is an incremental approach to build a data warehouse. Here we build the data marts separately at different points of time as and when the specific subject area requirements are clear. The data marts are integrated or combined together to form a data warehouse. Separate data marts are combined through the use of conformed dimensions and conformed facts. A conformed dimension and a conformed fact is one that can be shared across data marts.

A Conformed dimension has consistent dimension keys, consistent attribute names and consistent values across separate data marts. The conformed dimension means exact same thing with every fact table it is joined.

A Conformed fact has the same definition of measures, same dimensions joined to it and at the same granularity across data marts.

The bottom up approach helps us incrementally build the warehouse by developing and integrating data marts as and when the requirements are clear. We don't have to wait for knowing the overall requirements of the warehouse.

We should implement the bottom up approach when

1. We have initial cost and time constraints.
2. The complete warehouse requirements are not clear. We have clarity to only one data mart.

The advantage of using the Bottom Up approach is that they do not require high initial costs and have a faster implementation time; hence the business can start using the marts much earlier as compared to the top-down approach.

The disadvantages of using the Bottom Up approach are that it stores data in the de normalized format; hence there would be high space usage for detailed data. We have a tendency of not keeping detailed data in this approach hence losing out on advantage of having detail data i.e. flexibility to easily cater to future requirements. Bottom up approach is more realistic but the complexity of the integration may become a serious obstacle.

Design steps

The following nine-step method is followed in the design of a data warehouse:

1. Choosing the subject matter
2. Deciding what a fact table represents
3. Identifying and conforming the dimensions
4. Choosing the facts
5. Storing pre calculations in the fact table
6. Rounding out the dimension table
7. Choosing the duration of the db
8. The need to track slowly changing dimensions
9. Deciding the query priorities and query models

Technical considerations

A number of technical issues are to be considered when designing a data warehouse environment. These issues include:

- The hardware platform that would house the data warehouse
- The DBMS that supports the warehouse data

- Data for testing of hypothesis, trends and patterns
- Predefined repeatable queries
- Ad hoc user specified queries
- Reporting and analysis data
- Complex queries with multiple joins, multi level sub queries and sophisticated search criteria

Data extraction, clean up, transformation and migration

A proper attention must be paid to data extraction which represents a success factor for a data warehouse architecture. When implementing data warehouse several the following selection criteria that affect the ability to transform, consolidate, integrate and repair the data should be considered:

- Timeliness of data delivery to the warehouse
- The tool must have the ability to identify the particular data and that can be read by conversion tool
- The tool must support flat files, indexed files since corporate data is still in this type
- The tool must have the capability to merge data from multiple data stores
- The tool should have specification interface to indicate the data to be extracted
- The tool should have the ability to read data from data dictionary
- The code generated by the tool should be completely maintainable
- The tool should permit the user to extract the required data
- The tool must have the facility to perform data type and character set translation
- The tool must have the capability to create summarization, aggregation and derivation of records
- The data warehouse database system must be able to perform loading data directly from these tools

Data placement strategies

– As a data warehouse grows, there are at least two options for data placement. One is to put some of the data in the data warehouse into another storage media.

– The second option is to distribute the data in the data warehouse across multiple servers.

User levels

The users of data warehouse data can be classified on the basis of their skill level in accessing the warehouse.

There are three classes of users:

Casual users: are most comfortable in retrieving info from warehouse in pre defined formats and running pre existing queries and reports. These users do not need tools that allow for building standard and ad hoc reports

Power Users: can use pre defined as well as user defined queries to create simple and ad hoc reports. These users can engage in drill down operations. These users may have the experience of using reporting and query tools.

Expert users: These users tend to create their own complex queries and perform standard analysis on the info they retrieve. These users have the knowledge about the use of query and report tools

Benefits of data warehousing

Data warehouse usage includes,

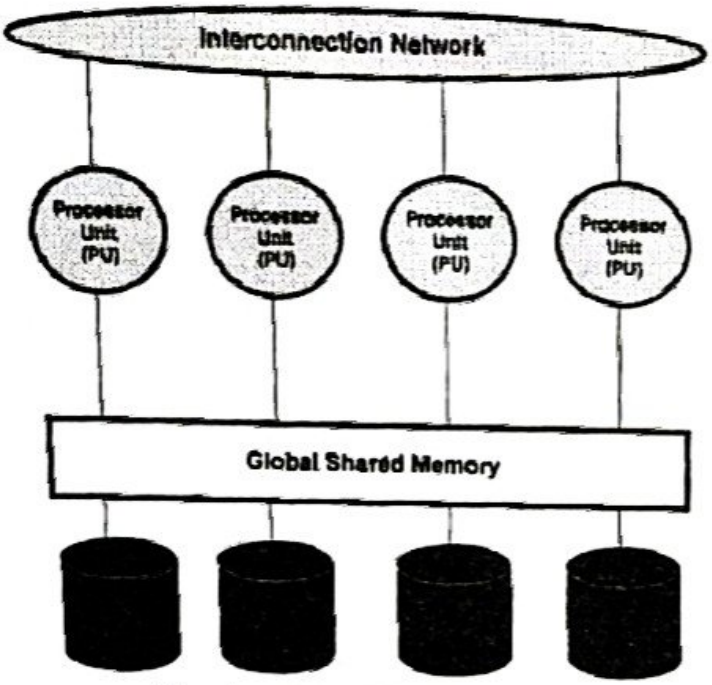
– Locating the right info

– Presentation of info

– Testing of hypothesis

– Discovery of info

– Sharing the analysis



Shared-memory architecture.

Parallel processing advantages of shared memory systems are these:

- Memory access is cheaper than inter-node communication. This means that internal synchronization is faster than using the Lock Manager.
- Shared memory systems are easier to administer than a cluster.

A disadvantage of shared memory systems for parallel processing is as follows:

- Scalability is limited by bus bandwidth and latency, and by available memory.

Shared Disk Architecture

Shared disk systems are typically loosely coupled. Such systems, illustrated in following figure, have the following characteristics:

- Each node consists of one or more PUs and associated memory.
- Memory is not shared between nodes.
- Communication occurs over a common high-speed bus.
- Each node has access to the same disks and other resources.
- A node can be an SMP if the hardware supports it.
- Bandwidth of the high-speed bus limits the number of nodes (scalability) of the system.

Shared nothing systems have advantages and disadvantages for parallel processing: Advantages

- Shared nothing systems provide for incremental growth.
- System growth is practically unlimited.
- MPPs are good for read-only databases and decision support applications.
- Failure is local: if one node fails, the others stay up.

Disadvantages

- More coordination is required.
- More overhead is required for a process working on a disk belonging to another node.
- If there is a heavy workload of updates or inserts, as in an online transaction processing system, it may be worthwhile to consider data-dependent routing to alleviate contention.

Shared disk architecture vs. shared nothing architecture	
Shared disk Architecture	Shared nothing architecture
Requires special hardware	Does not require special hardware
Non linear scalability	Provides near linear scalability
Balanced CPU or node fail-over	Balanced/Unbalanced CPU or node fail-over
Requires CPU level communication at disk access	Minimal communication
Non disruptive maintenance	Non disruptive maintenance

Parallel DBMS features

- Scope and techniques of parallel DBMS operations
- Optimizer implementation
- Application transparency
- Parallel environment which allows the DBMS server to take full advantage of the existing

DBMS schemas for decision support

The basic concepts of dimensional modeling are: facts, dimensions and measures. A fact is a collection of related data items, consisting of measures and context data. It typically represents business items or business transactions. A dimension is a collection of data that describe one business dimension. Dimensions determine the contextual background for the facts; they are the parameters over which we want to perform OLAP. A measure is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions. Considering Relational context, there are three basic schemas that are used in dimensional modeling:

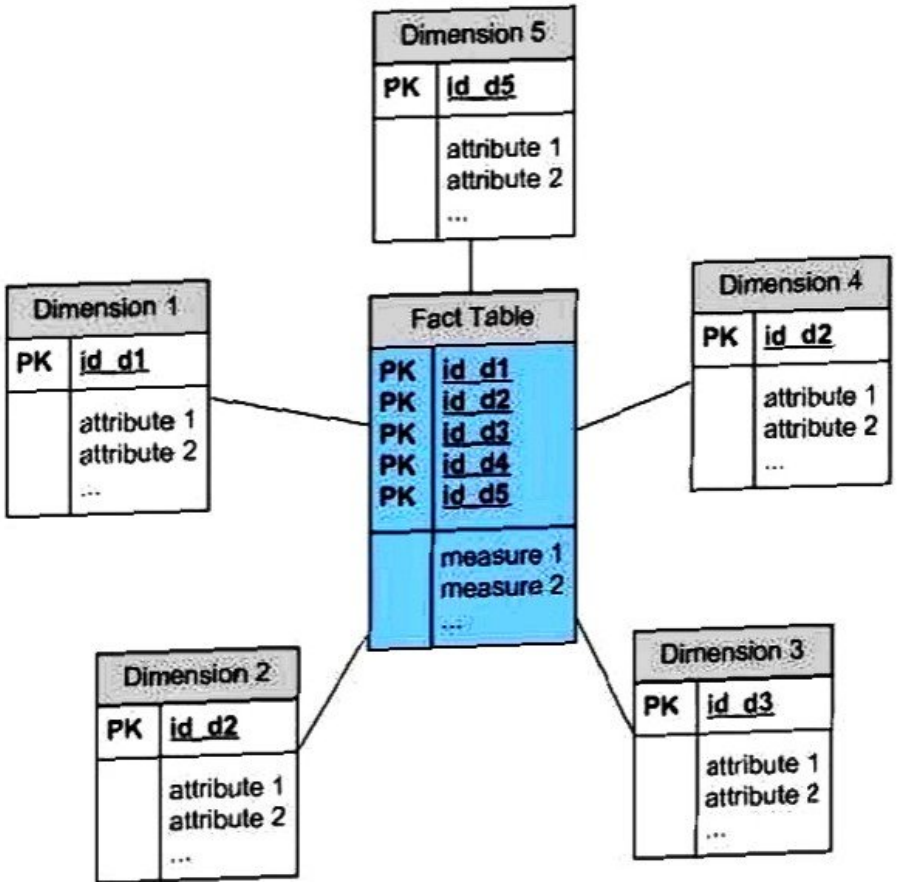
Star schema

The multidimensional view of data that is expressed using relational data base semantics is provided by the data base schema design called star schema. The basic of stat schema is that information can be classified into two groups:

- Facts
- Dimension

Star schema has one large central table (fact table) and a set of smaller tables (dimensions) arranged in a radial pattern around the central table.

Facts are core data element being analyzed while dimensions are attributes about the facts. The determination of which schema model should be used for a data warehouse should be based upon the analysis of project requirements, accessible tools and project team preferences.



The star schema architecture is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of fact table and the points of the star are the dimension tables. Usually the fact tables in a star schema are in third normal form(3NF) whereas dimensional tables are de-normalized. Despite the fact that the star schema is the simplest architecture, it is most commonly used nowadays and is recommended by Oracle.

Fact Tables

A fact table is a table that contains summarized numerical and historical data (facts) and a multipart index composed of foreign keys from the primary keys of related dimension tables. A fact table typically has two types of columns: foreign keys to dimension tables and measures those that contain numeric facts. A fact table can contain fact's data on detail or aggregated level.

Dimension Tables

Dimensions are categories by which summarized data can be viewed. E.g. a profit summary in a fact table can be viewed by a Time dimension (profit by month, quarter, year), Region dimension
(profit by country, state, city), Product dimension (profit for product1, product2). A dimension is a structure usually composed of one or more hierarchies that categorizes data. If a dimension hasn't got a hierarchies and levels it is called flat dimension or list. The primary keys of each of the dimension tables are part of the composite primary key of the fact table.

Dimensional attributes help to describe the dimensional value. They are normally descriptive, textual values. Dimension tables are generally small in size then fact table. Typical fact tables store data about sales while dimension tables data about geographic region (markets, cities), clients, products, times, channels.

Measures

Measures are numeric data based on columns in a fact table. They are the primary data which end users are interested in. E.g. a sales fact table may contain a profit measure which

Data Extraction, Cleanup and Transformation Tools

- The task of capturing data from a source data system, cleaning and transforming it and then loading the results into a target data system can be carried out either by separate products, or by a single integrated solution. More contemporary integrated solutions can fall into one of the categories described below:
  - Code Generators
  - Database data Replications
  - Rule-driven Dynamic Transformation Engines (Data Mart Builders)

Code Generator:

- It creates 3GL/4GL transformation programs based on source and target data definitions, and data transformation and enhancement rules defined by the developer.
- This approach reduces the need for an organization to write its own data capture, transformation, and load programs. These products employ DML Statements to capture a set of the data from source system.
- These are used for data conversion projects, and for building an enterprise-wide data warehouse, when there is a significant amount of data transformation to be done involving a variety of different flat files, non-relational, and relational data sources.

Database Data Replication Tools:

- These tools employ database triggers or a recovery log to capture changes to a single data source on one system and apply the changes to a copy of the data source data located on a different system.
- Most replication products do not support the capture of changes to non-relational files and databases, and often do not provide facilities for significant data transformation and enhancement.
- These point-to-point tools are used for disaster recovery and to build an operational data store, a data warehouse, or a data mart when the number of data sources