Unit I
Data Warehousing

Introduction to Data Warehousing

- Definition based on Computing
  - * Data Warehouse is also known as Enterprise data Warehouse,
  - * it is used for reporting and data analysis and Considered a Core Component of buscines intelligence.
  - * It is a central responsibores [place where data restored] of integrated data from one or more disparate Sources.
  - * It store current and historical data in one Single place and used for Creating analytical reports for knowledge workers.

The Need of Data Warehousing.

In Buscines

- * Decision need to be made quickly and Correctly using all available data
- * Users are buscines domain experts, not Computer Professionals
- * The amount of data doubles every 18 months which affects the response time
- * Competition in the areas of buscines intelligence is high with added adoled Information.

Technology reason for the existence of DW

- The DW is designed to address the incompatibility of informational and operational transactional systems.
- The IT infrastructure is changing rapidly and capabilities are increasing as follows
  - The price of computer processing speed in MIPS [Million instructions per seconds] continues to decline whereas the power of Micro processor doubles every 2 years
  - The price of digital storage is rapidly dropping
  - N/w bandwidth is increasing, while the price of high bandwidth is decreasing
  - Workplace is increasing in terms of S/w & H/w
  - Legacy systems need to and can be integrated with new applications

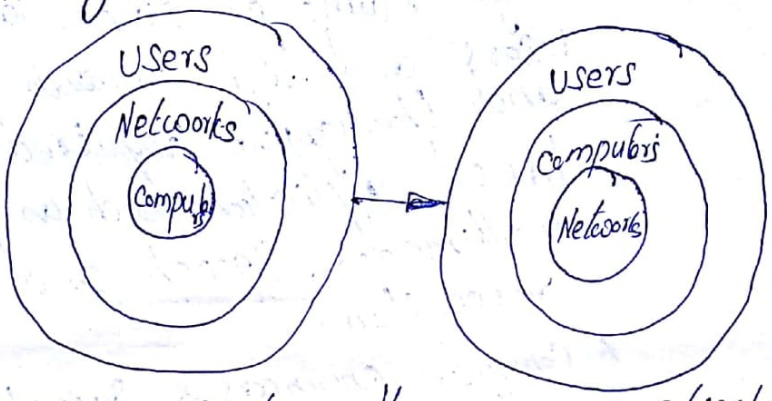
Paradigm Shift:

- Based on client Server Architecture.

(a) Computing Paradigm
(b) Business paradigm

Computing paradigm:

- Computer user that access a powerful tool - the Computer via a Communication N/w.
- User use Computers to solve problems and to request services.
- users use their individual computers to serve as prob entry points to get access to this distributed Computing power



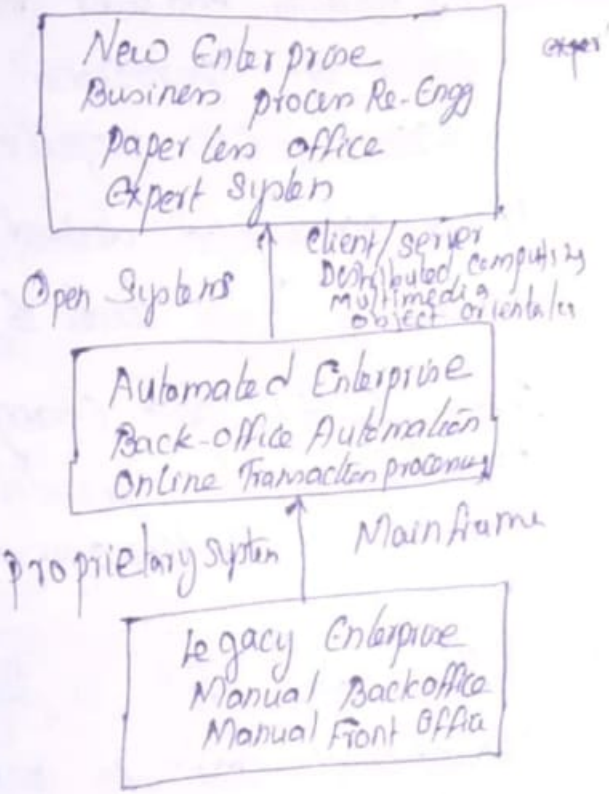
- The major changes affecting the way client/server Computing is implemented do not invalidate the Clinent/Server architecture, but introducing some requirements on the environment. They are
  - Object Orientation: Productivity of the it development, Object Oriented analysis design and programming cannot be ignored any longer
  - Middleware: The layer in the client/server architecture that transforms a simple two-tier client/server Computing model into a more complex client-middleware-Server model

* Storage and handling of Complex data:
  - the ability to store and manipulate Complex types of data such as video, image text and Spatial and time Series data in new-relational DB
- High-performance Commercial parallel Computing and Very large database [VLDB] processing:
  - Servers are incapable of handling large Volume of data, large number of users and high demands on performance and throughput required of the new breed application such as datamining, multimedia, Speech and character recognition.
  - Computer engineers & Scientists give their attention to different Comp. Architecture and database processing to satisfy these demands

(ii) Business paradigm.
- Computing technology change the way we do Business
- proprietary ownership System & mainframe-based Computing in 60s & 70s move the traditional business enterprise from manual back and front office to an automated back office & on-line Transaction processing (OLTP.

* Automated Systems were designed to duplicate manual operations, but it improves the processing speed and throughput.

The role of information technology



~~Business problem definition.~~

Operational and Informational Data Stores

* Operational data focuses on transactional functions such as bank card withdrawals and deposits
* Informational data, on the other hand is organized around subjects such as customer, vendor and product
* Informational data is obtained from operational data sources.
* Both uses relational database management System

Characteristics differ of Informational data difference from operational data

Data accen : it tends to be adhoc, rather than Predefined Structured accen

Data Model (Schema) : It reflects end user analysis need, while an operational data model is while support ACID properties

Time base : Recent, aggregated, derived and historical data. Operational data tends to be current data

Data changes : Informational data changes are mostly periodic, while operational data is subject to continuous high-frequency changes

Unit of work : Informational data is Queried, while Operational data is subject to Concurrent update, insert and delete.

Records range accened per transaction : Millions for informational data versus tens for operational

Number of concurrent users : hundreds for Informational versus thousands for Operational

Transaction volume : Relatively low for informational data but high for operational data.

Types of users : Analytical, Managerial vs. Clerical; Operational Users; user of the operational data is another system

Difference b/w informational and operational data base.

	Operational data	Informational data
Data Content	Current values	Summarized, archived, derived
Data Organization	By application	By subject
Data Stability	Dynamic	Static
Data Structure	optimized for transaction	optimized for complex queries
Access Frequency	High	Medium to low
Access Type	Read/update/delete	Read/aggregate
Response time	Subsecond (<1sec)	Several seconds to minutes

Operational data Store;

- ODS is an architectural Concept. to Support day-to-day operational decision support and current value data propogated from operational application
- ODS provides an alternative to Operation DSS application accessing data directly from OLTP systems

Significant challenges of ODS are

- Location of the appropriate source of data
- Transformation of the source data to satisfy the ODS data model
- Complexity of near-real-time propogation of changes from the operational system to the ODS
- DBMS combines effective query processing with transactional processing capabilities [ACID]

Datawarehouse Definition & Characteristics

It is viewed as an information system with the following attributes.

- It is a database designed for analytical tasks. using data from multiple application
- Small number of users with long interactions
- Its usage is read-intensive
- Its content is periodically updated
- Contains current and historical data
- Contains a few large tables

Definition based on Inmon:

"A datawarehouse is a subject-oriented, integrated, time variant, non volatile collection of data.

Terms related to Datawarehouse

Current detail Data:

- Data is acquired directly from the operational database.
- The current detail data include customer profile data, customer activity data.

Old detail Data

- This represents historical data or aged current detail data

Data Mart:

- An implementation of the DW is called. Its data scope is limited. Compared to the enterprise wide DW

Summarized data:

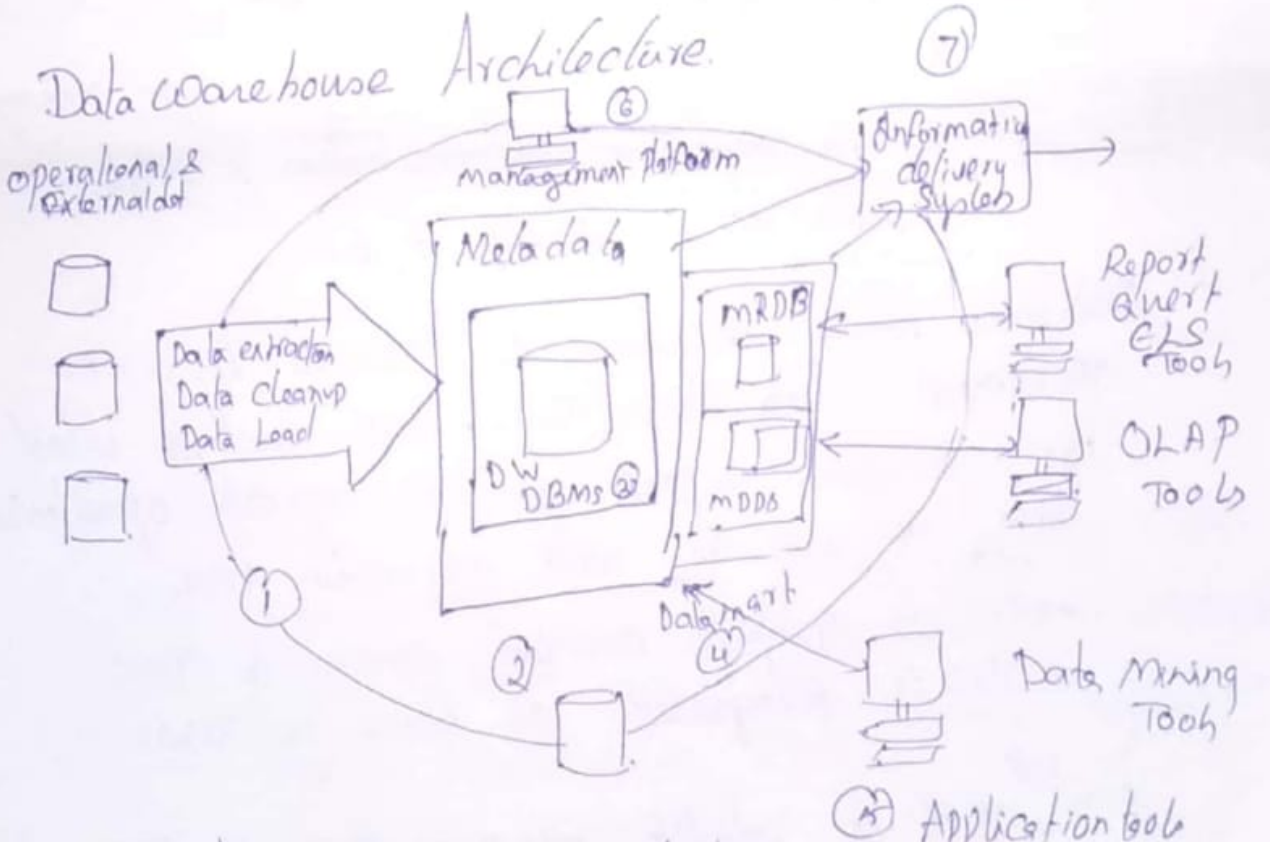
- Data is aggregated along the lines required for executive-level reporting, trend analysis and enterprise wide decision making

Drill down

- An ability of a knowledge worker to perform business analysis in a top-down fashion

Meta data: data about data, which contains location and description of DW system components, name, definition, structure etc.

Data warehouse Architecture



- Data warehouse architecture is based on relational database mgt server that fns as a central respository for informational data

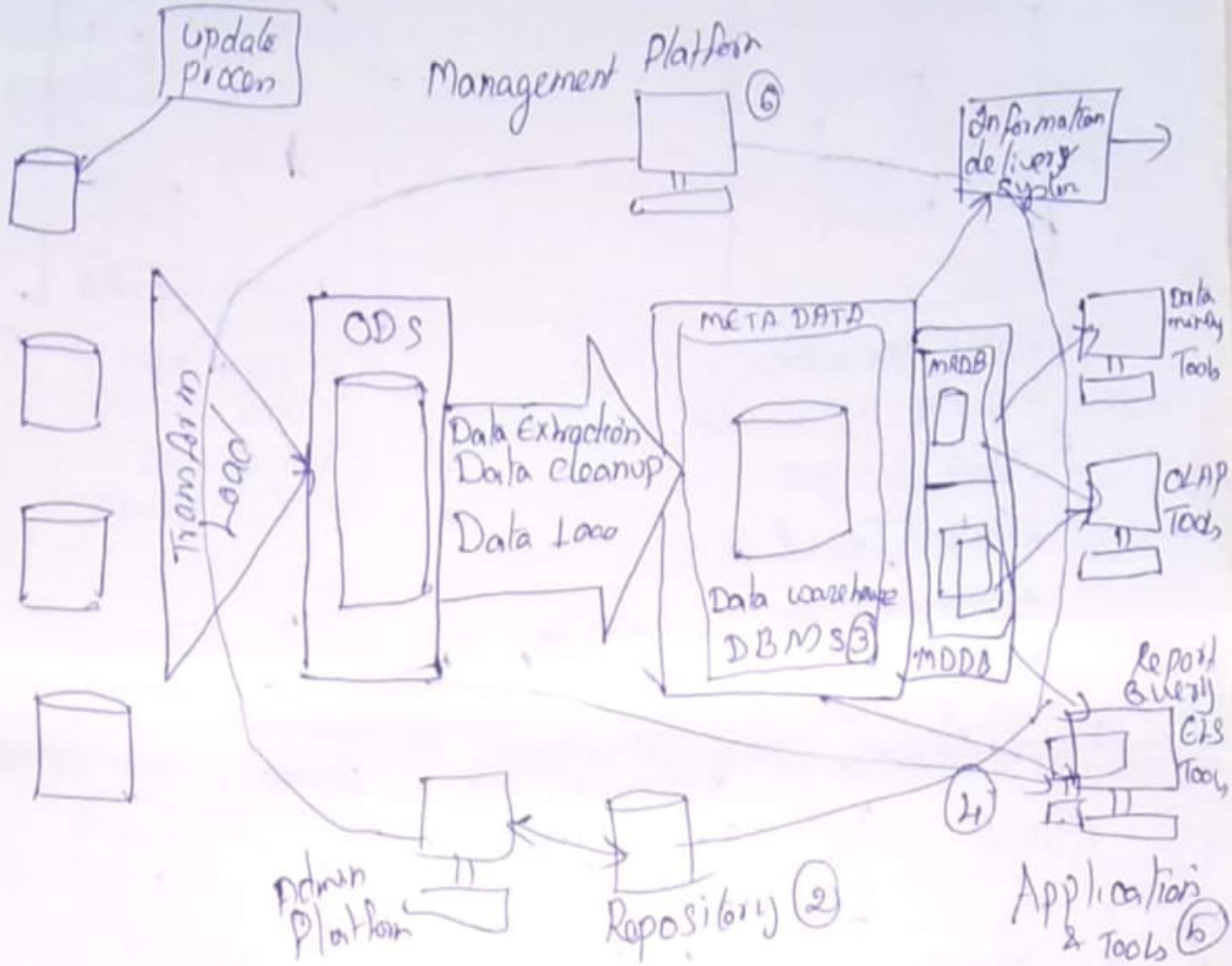
* The source data for the coarehouse is the operational applications
* Once the data enters the DW, it is transformed into an integrated structure and format.
* transformation process involve conversion, summarization, filtering & condensation of data.
* Data coarehouse contains large historical component
* DW Architecture consist of 7 components
  * Data sourcing, cleanup, transformation and migration tools
  * Metadata resposibry
  * coarehouse database technology
  * Data Marts
  * Data Query, reporting analysis and mining tools
  * Data coarehouse administration & Mgt
  * Information delivery system

Data coare house architure is extended by using operational data stores [ODS]. ODS can be used for decision support activies against operational data as coell as data acquisition area.

Simile
* ODS is subject-oriented similar to DW
* ODS is integrated as same as DW.

Diff
* ODS is volatile where as DW is non-volatile
* ODS contain current data but DW contains Both

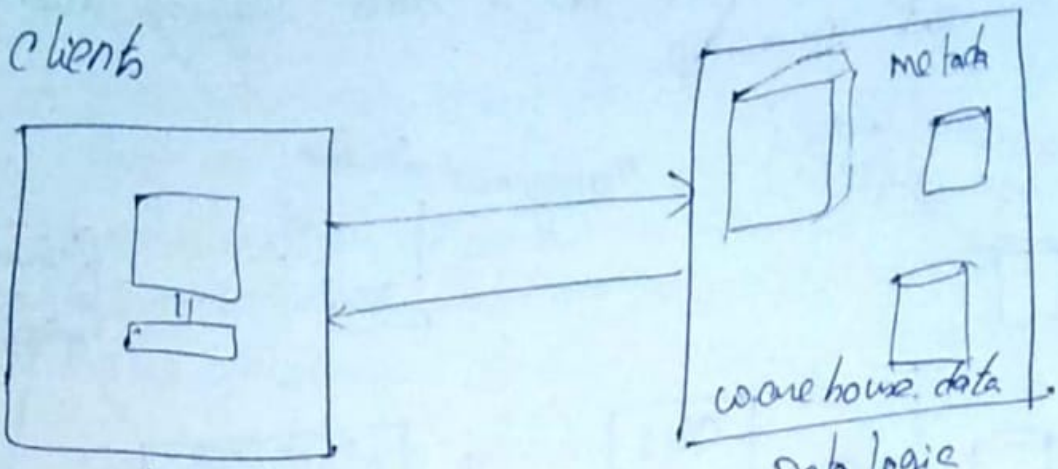
ODS may be used as a data staging area for DW data sourcing.



Two tiered Architecture

- is a 'fat' client model or colock client system function includes user interface Query Specification, data Analysis, report formatting, aggregation, & data access.
- Server performs data logic, data services, file services and maintains ~~da~~ metada

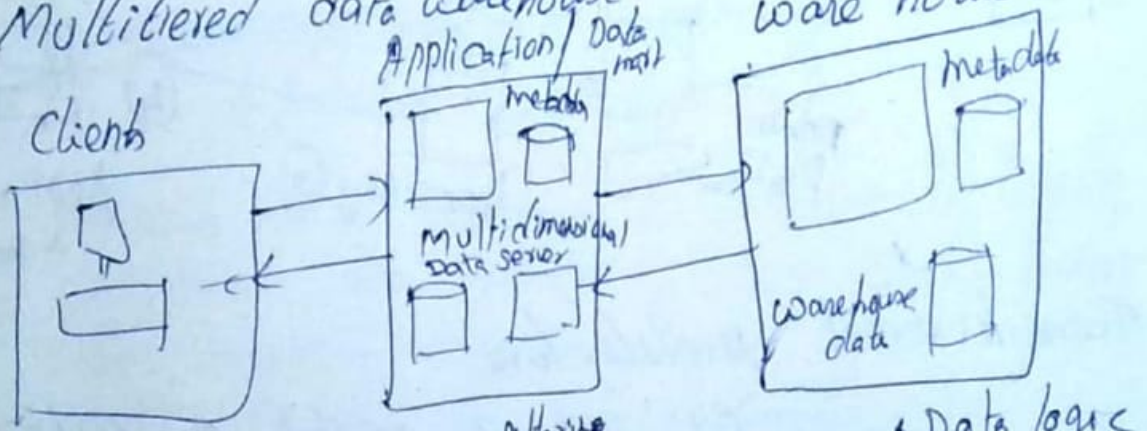
- it lacks scalability & flexibility



- GUI / presentation
- Query specification
- Data Analysis
- Report formatting
- summarizing
- Data access

- Data logic
- Data services
- Meta data
- File services

Multitiered data warehouse architecture,



- GUI/presentation,
- Query Specification
- Data Analysis
- Report formatting
- Data Access

- Filtering
- Summarizing
- Meta data
- Multidimensionalities
  Data access

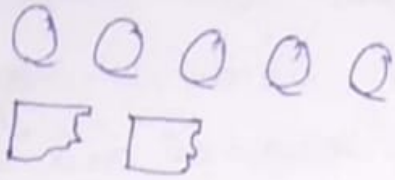
- Data logic
- Data services
- meta data
- file service

- It solves the scalability & flexibility issues of two tier architecture

Evolution of Decision Support System

- Origin of datawarehousing and decision support system (DSS) processing to the very early days of computers and information system

1960



Master files, reports

1965

Lots of Master files

- Complexity of-
  - Maintenance
  - Development
- Synchronization of Data
- Hardware

1970

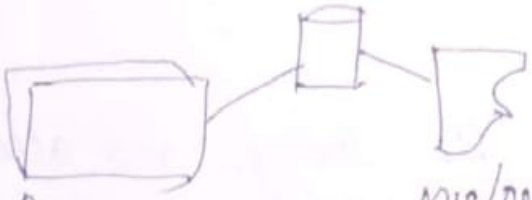
DASD
DBMS

Database – "a single source of data for all processing

1975

Online, high-performance transaction processing

1980.



Tx Processing

MIS/DSS

PCs, 4GL technology

1985

The Single database-serving-all purpose paradigm

* 1960's the world of Computation consist of Creating individual application that runs using Master files
* Application featured reports & and programs using early language such as COBOL or fortran.
* punched Cards and paper tape were common.
* Magnetic to Master files are Stored in Magnetic tape
* Magnetic tape Store large volume of data, cheap. drawback the data are accessed Sequentially

* 1965's or Mid-1960's
  * the growth of Masterfiles and magnetic tape exploded.
  * the Master files and redundant data presented some problem
    * The need to Synchronize data upon update
    * the Complexity of Maintaing pgm
    * the Complexity of developing pgms
    * the need for extensive amounts of h/w to support all the Master file

* 1970's
  * A new technology for Storage and access of data had been used
  * disk storage or direct access storage device [DASD]
  * Disk storage is differ from magnetic tape. the data Can be accessed directly on DASD.
    * to get the value of some records it is necessary to know the address of record so that it can be access directly

- with DASD new type of system software came ie DBMS
- DBMS make easy for the programmer to store and access data on DASD.

1970's Online transaction procesing (OLTP) for ex.

1980 PC & 4 GL (fourth generation Languages) has began
- MIS is used for to drive Management

~~1985~~
1985 Enter the Extract Program
- The Extract program is the simplest of all pgms.
- It becomes popular for two reasons
  (i) Extract procesing can move data out of the way of high performance Online procesing, there is no Conflict in terms of performance
  (ii) when data is moved out of the operational, transaction-procesing domain with an extract program, a Shift in Control occurs

Extract program
- performance
- Control

→ start with some parameters, search a file based on the satisfaction of the parameters

Scanned by CamScanner

Problems with Naturally Evolving Architecture.

- Data Credibility
- Productivity
- Inability to transform data into information

Lack of Data Credibility

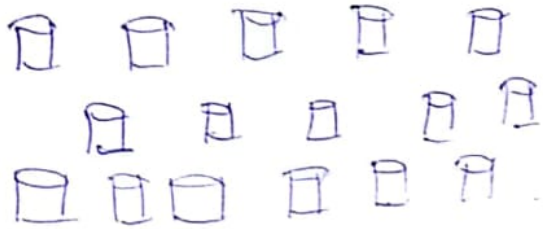
Five Reasons

- No time basis of data
- The algorithmic differential of data.
- The problem of external data
- Levels of extraction
- No common source of data from the begining

Problems with productivity

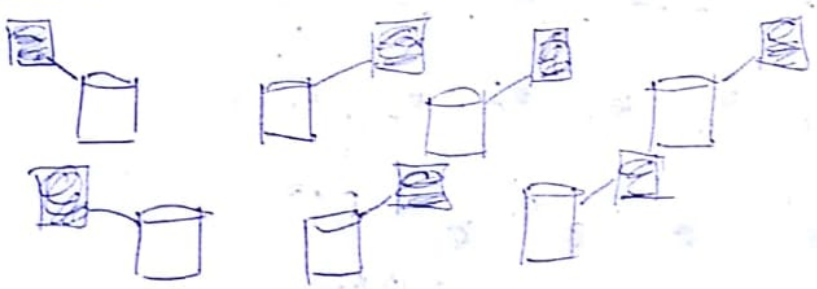
- Many files and collections of data, - how to create correct report for an organisation
  - Locate and analyze the data for report
  - Compile the data for the report
  - Get programmer/ analyst resource to accomplish these two task

Productivity



Produce a corporate report, access all data

Locating the data requires looking at lot of files

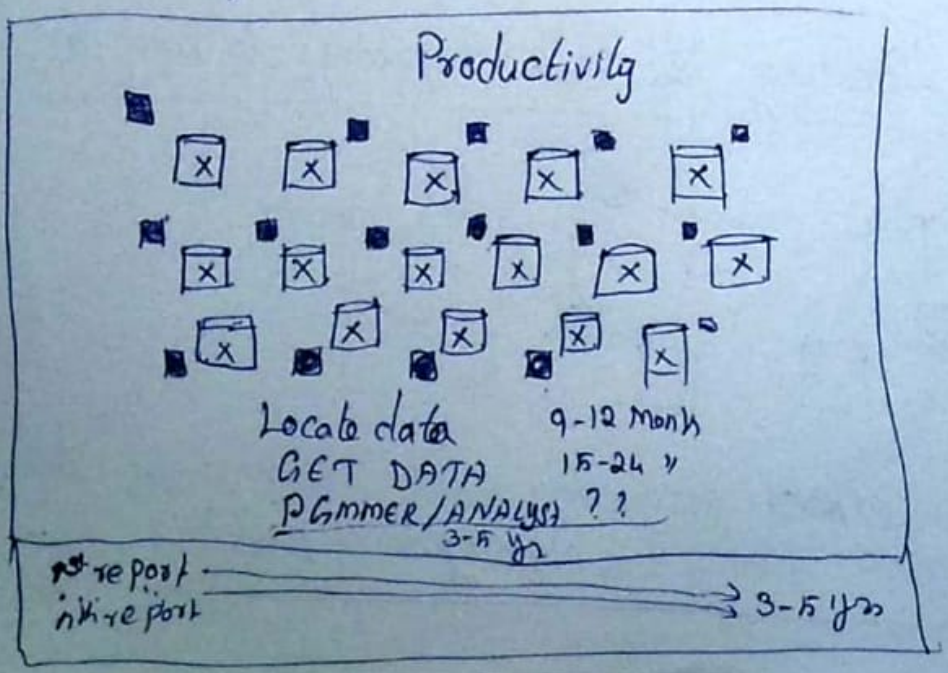


Lots of extract program, each customized, have to cross many technological barriers

(i) * to locate the data many files and layouts must be analyzed.
* Some files uses the Virtual storage Access Method (VSAM), some files uses Information Management System, some uses Integrated Database Management System (IDMS).
* Different skills sets are required in order to access data across the enterprise.

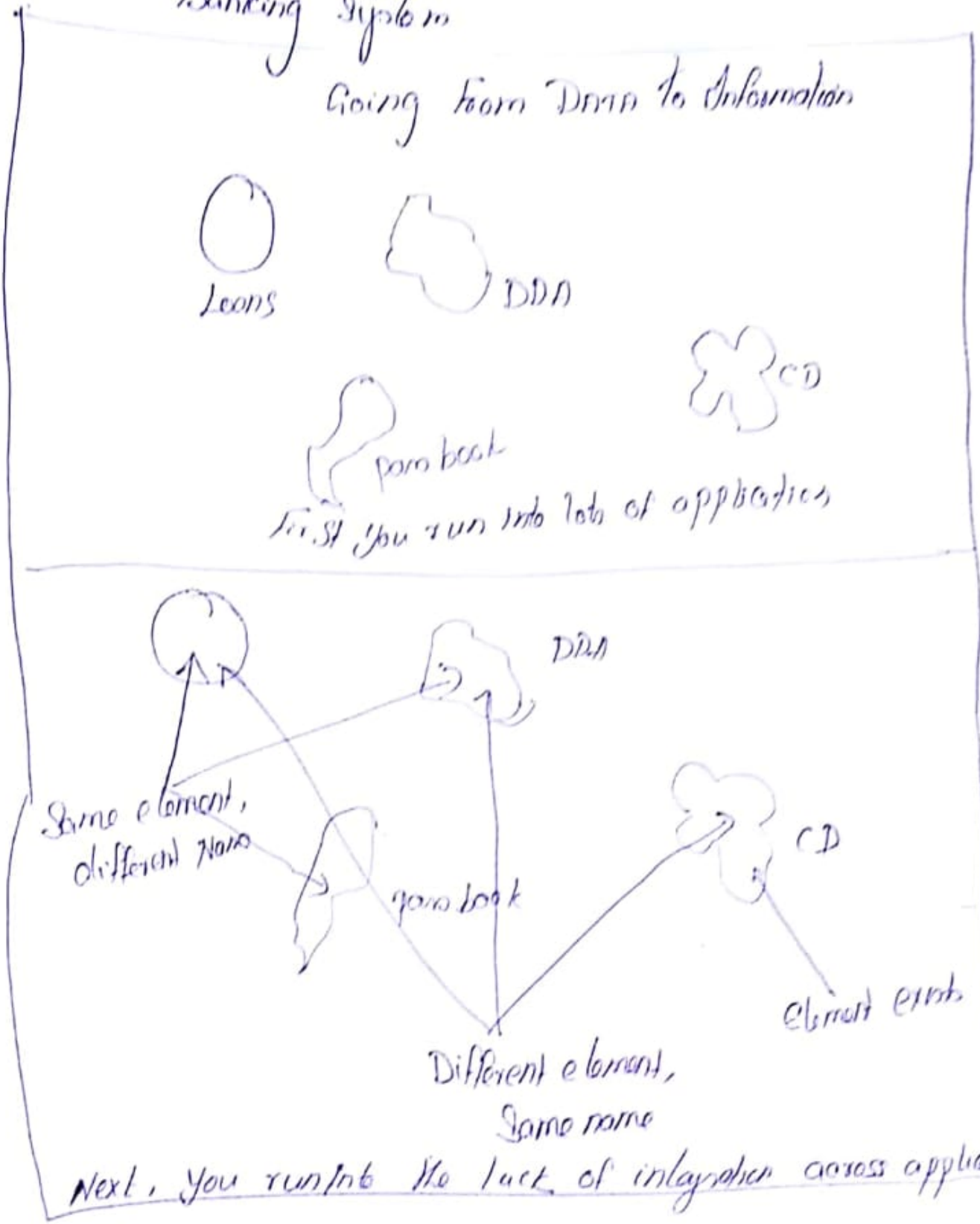
(ii) Producing the report is to compile the data once it is located. Complications facts are
* Lots of programs have been written
* Each program must be customize
* The program cross every technology that the company use

(iii) * If the designer has asked for one or three months of resource then generating a report might not have required much attention.
* If the analyst request for many resources then the request must be considered



From Data to Information

- The major fault in the evolving architecture is the inability to go from data to information
- Banking System



① First thing

- The DSS analyst discovers in trying to satisfy the request for information is that going to existing system.
- DSS analyst have to deal with lots of unintegrated non-integrated legacy applications

② There is not enough historical data stored in the applications to meet the needs of the DSS request

A change in Approach

- Two types of data in the architected environment
  - primitive data
  - derived data

Primitive data/ operational data	Derived data/DSS data
Application oriented	Subject oriented
Detailed	Summarized
Accurate	Represents values over time
Can be updated	Is Not updated
Accessed a unit at a time	Accessed a set at a time
Transaction driven	Analysis driven
Highly availability	Related availability
Non redundancy	Redundancy
Static Structure	Flexible Structure
High probability of access	Low probability of access

The Architected Environment

* four levels of data on Architected Environment
  (i) operational level.
  (ii) atomic or data warehouse level
  (iii) departmental or data mart level
  (iv) Individual level.
* These levels of data are the basis of larger architecture called Corporate Information factory (CIF)
* Operational level holds application-oriented primitive data only and serves the high performance transaction processing community
* DW level hold integrated, historical primitive data that cannot be updated
* Data mart level contains derived data. it is shaped by the end user requirements
* Individual level which has heuristic data Analysis

Level of Architecture

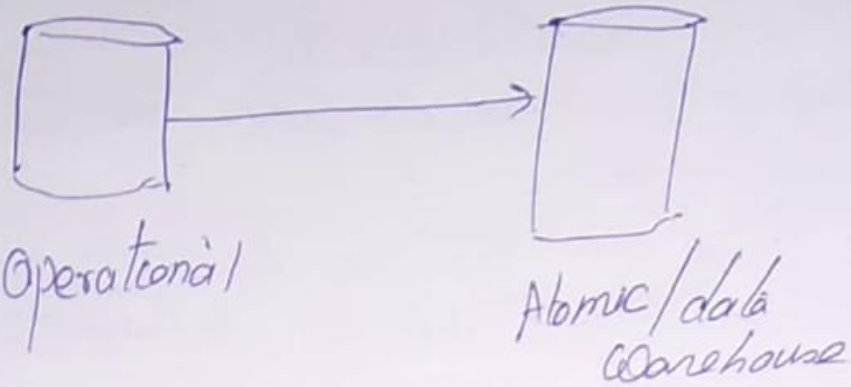
* Detailed
* Day-to-day
* Consistent valued
* Highly probability of accessing
* Application oriented

* Time variant
* integrated
* Subject-oriented

* derived & Primitive
* Departmental
  * Accounts
  * Marketing
  * Engineering

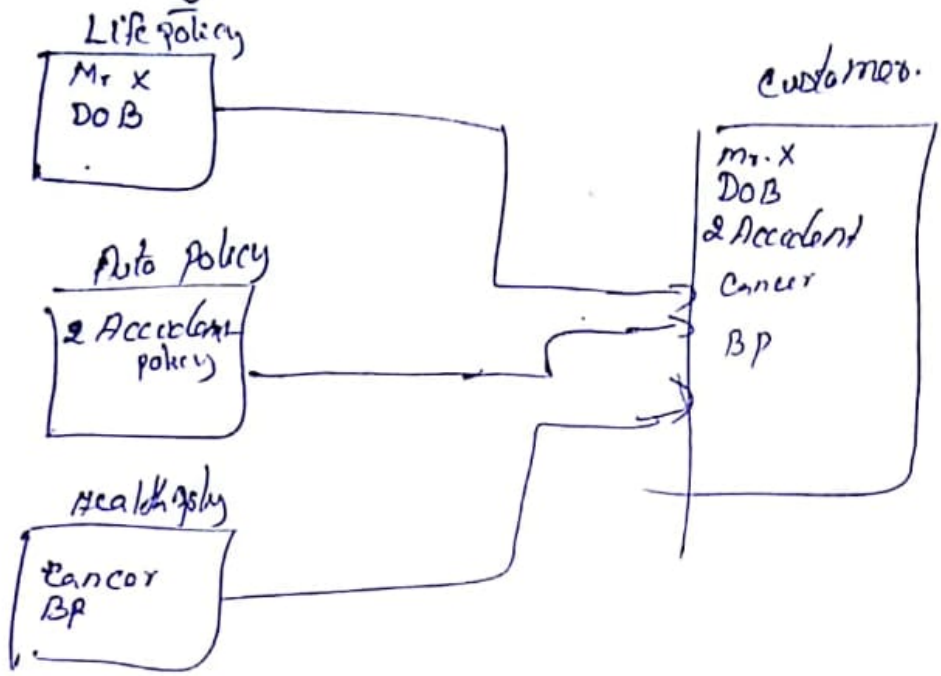
* Temporary
* Adhoc
* Heuristic
* Non-repetitive

Data integration in the Architected Environment



- As the data passes from operational environment to the data warehouse environment it is integrated
- if the data arrives at the data warehouse without integrating, it cannot be used to support Corporate view of data
- unintegrated data is complex and difficult to use.
- Extract/transform/Load (ETL) s/w can be used to automate the process

eg



ED

Granularity in the Data coare house

- Determining the proper level of granularity of the data that coill reside in the data coare house
- it is important to the coare house architect because it affect all the environments that depend on the coare house

Granularity :

- is the level of depth represented by the data or dimensional table in a data coarehouse.

Raw Estimate:

- number of rows of data that coill reside in the data coarehouse tells the architect a great deal

~~Input to the planning~~

Estimating row/space for the coarehouse Environment

1. for each known table:
   How big is a row (in bytes)
   - biggest estimate
   - Smallest estimate

For the 1-year horizon
what is the maximum number of rows possible?
what are the minimum number of rows possible?

For the 2-year horizon.
what is the maximum number of rows possible?
what is the minimum number of rows possible?

For each key of the table.
what is the size of the key (in bytes)?

Total maximum 1-year space = biggest row × 1-year max rows

Total minimum 1-year space = smallest row × 1-year min rows

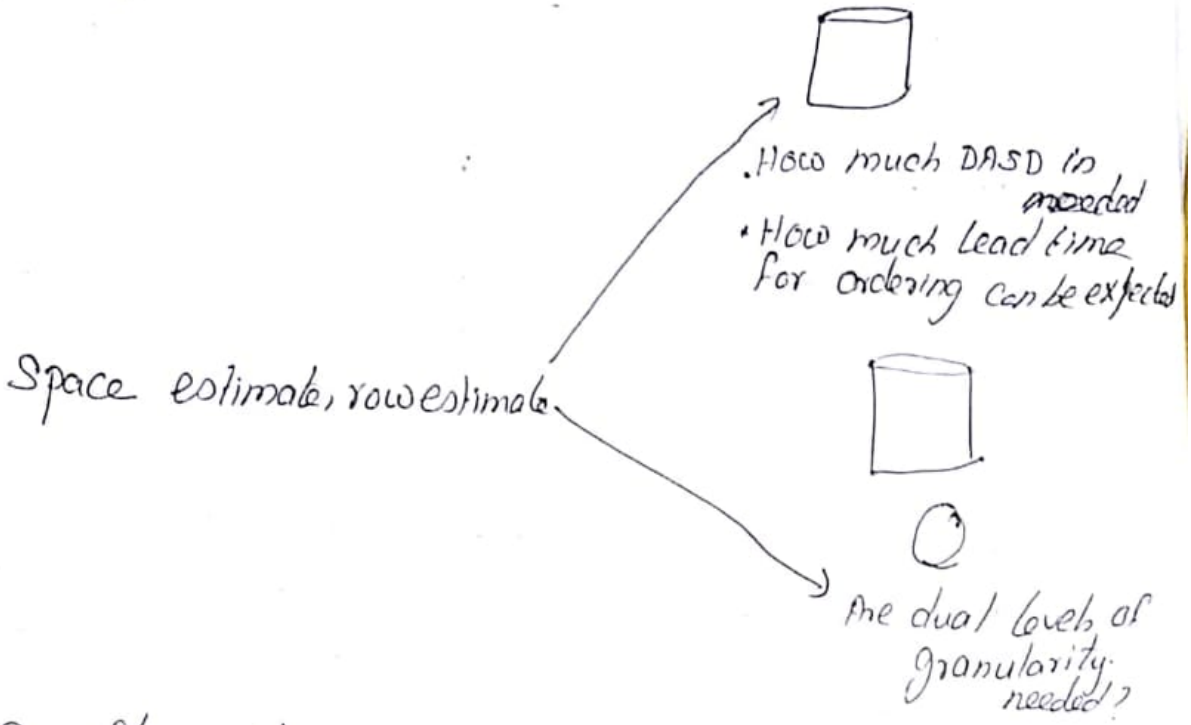
2. Repeat (1) for all known tables

The Algorithm Calculate the space occupied by a data warehouse.

- 1st step identify all the tables to be built.
- As the rule of thumb, there will be one or more large tables and many small supporting tables.
- Next estimate the size of the row in each table with lower bound and upper-bound estimate
- Next estimate one year horizon

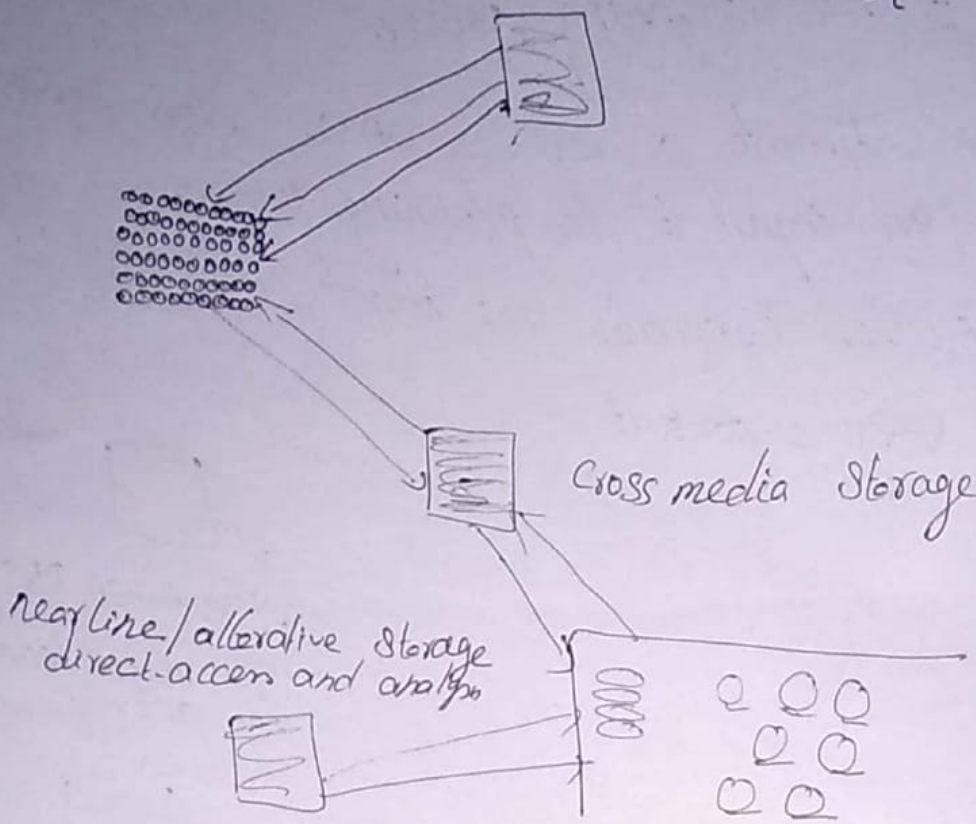
Input to the planning process

* Estimate of rows & DASD then serves as input to the planning process.
* When estimates are made, accuracy is very important

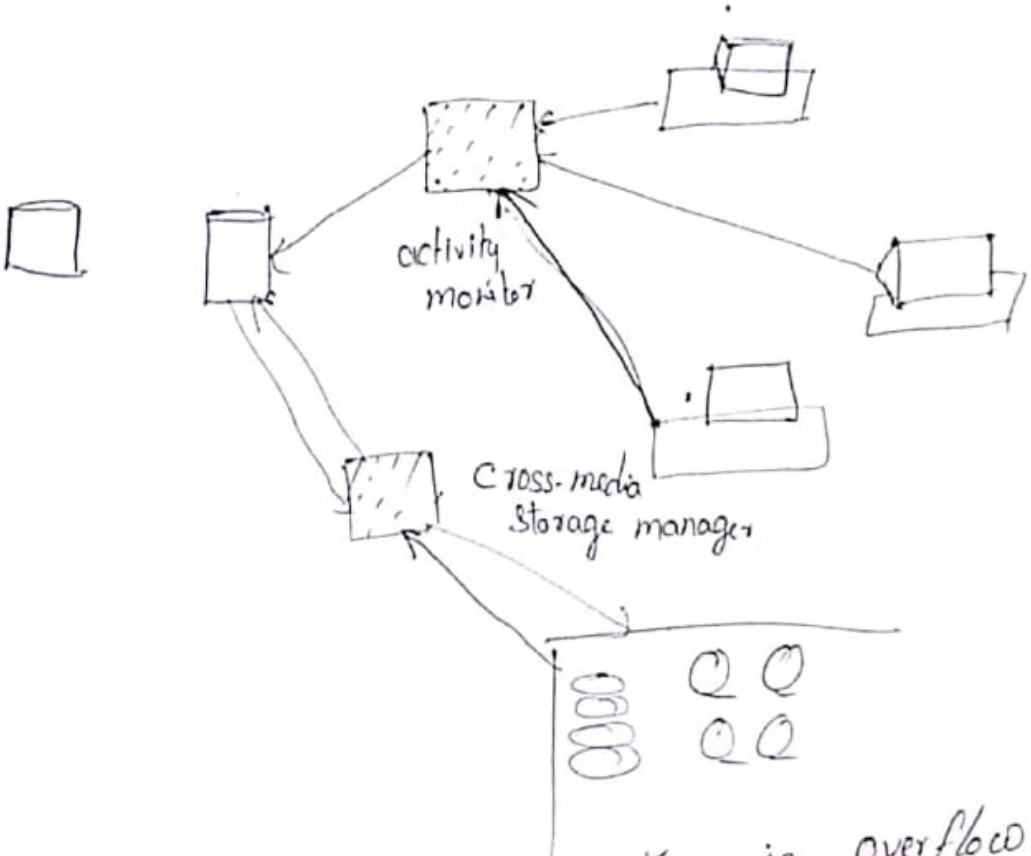


Overflow Storage

* Data in the Data warehouse grows high with the combination of historical and detailed data.
* It is mentioned as tera bytes & peta bytes.
* Datas are subdivided into active and inactive data.



- Data Monitor determines the usage of data.
  - It tells where the data has to be placed.
  - The Movement between the disk storage and near-line storage is controlled by the software called Cross media storage Management.
  - The data in the alternative storage can be accessed by the S/w, which identifies where the data is stored in the near line.
  - In the overflow storage, not frequently used data are stored
  - The overflow storage can be any number of storage media they are, photo optical storage, Magnetic tape, and cheap disk.



- Two pieces of S/w are there in overflow storage
  (i) Cross media storage mange.
  (ii) activity monitor

Cross

- manages the traffic of data going to and from the disk storage environment to alternate storage environment
- Data is moved ~~where~~ based on the request
- because of this administrator is able to get maximum performance from the system

(ii) Activity Monitor

- Determines what data is and is not being accessed.
- It determine where data is placed whether in disk storage or alternate storage

Building a Data warehouse

Busines Considerations: Return on Investment

(i) Approach

Two approaches

(i) top-down approach
(ii) Bottom-up approach

Top-down approach

* An organization has developed an enterprise data model, collected enterprise wide busines requirement and decided to build an enterprise data warehouse with subset data marts

Bot-up approach

* Busines priorities resulted in developing individual data marts, which are integrated into the enterprise data warehouse
* It is more realistic

(ii) Organizational issues:

* organization it is expert in developing operational system
* The requirement & environments associated with the information application of data warehouse.
* ∴ the organization has to use different development practices than the ones it uses for operational applications

Design Considerations

The main factors in design includes

- Heterogeneity of data source, which affects data conversion, quality & timeliness
- use of historical data
- Tendency of database to grow very large
- data warehouse design is different from traditional OLTP
- DW is business driven it requires continuous interaction with end users

(i) Data Content:

- Content & structure of data warehouse are reflected in its data model
- The data model is a template that describes how information will be organized within the integrated warehouse framework
- It identifies major subjects and relationships of the model including keys, attributes and attribute grouping

(ii) Metadata

- it defines the contents and location of data in the warehouse, relationships between the operational data bases and the data warehouse

- Metadata provides decision support oriented pointers to warehouse data, provides logical link between warehouse data and the decision support applications.

(iii) Data distribution:
- data volume continue to grow,
- how data should be divided across multiple servers and which user should get access to which type of data.
- data distribution to based on subject area, location, time

(iv) Tools
- provide facilities for defining the transformation and cleanup rules, data movement, end user query, reporting and data analysis.

(v) Nine decision in the design of data warehouse
(i) Choosing the subject matter
(ii) Deciding what a fact table represent [large central table]
(iii) identifying and conforming the dimensions
(iv) choosing the facts
(v) Storing precalculations in the fact table
(vi) Rounding off the dimensions table
(vii) Choosing the duration of the database
(viii) The need to track slowly changing dimension
(ix) Deciding the query priorities & query modes

Technical Consideration

Issues include

Sorting
Join
Summarization
Formatting

* The h/w platform that could house the data warehouse [capacity for handling decision support applications]
* The DBMS that supports the warehouse database
* The Communications infrastructure that connects the warehouse, datamarts, operational systems and end user
* The hardware platform and software to support the meta data repository
* The System management framework that enables centralized management and administration of the entire environment

Implementation Considerations

The Complexity to Implement or build the data ware house are as follow

* Collect & Analyze business requirement
* Create data model and a physical design for the data warehouse

- Define data sources
- Choose database technology and platform for the coarehouse
- Extract the data from the operational database, transform it, clean it up and load it into the data base
- Choose database accen and reporting tools.
- Choose database Connectivity software
- Choose data analysis and presentation software
- Update the data coarehouse

(i) Accen tools

- Accen type include
  - Simple tabular form reporting
  - Ranking
  - Multivariable analysis
  - Time Series analysis
  - Data visualization, graphing, charting and pivoting
  - Complex textual search
  - Information mapping
  - Adhoc user-specified queries
  - predefined repeatable queries
  - Interactive drill-down reporting and analysis

(ii) Data extraction, cleanup, transformation and migration.

* The ability to identify data in the data source environments that can be read by the conversion tool.
* Support for flat files, indexed files, and legacy DBMS.
* Capability to merge data from multiple data stores
* The ability to read information from data dictionaries from resposibory
* The code generated by the tool should be completely maintainable from within the development environment
* The ability to perform data type and character set translation
* Capability to create summarization, aggregation and derivation of records
* able to perform loading data directly from these tools

(iii) User Sophistication levels:

Three classes of users

* Casual users
* power users
* Expert users

Casual users:

- are most comfortable retrieving information from the warehouse in predefined format and preexisting queries and report

Power users:

- Combine predefined quires with some simple adhoc queries created by themselves

Expert users:

- Create their own complex queries to perform standard analysis on the information they retrieve from the warehouse.

Benefits of Data warehousing.

- Locating the right information
- presentation of information (reports, graphs)
- Testing of hypothesis
- Discovery of information
- Sharing the analysis

Tangi benefits

- product inventory turnover is improved
- Cost of product are decreased with improved selection of target markets

- Better business intelligence is enabled by increased quality and flexibility of market analysis

Intangible benefits

- Improved productivity, by keeping all required data in a single location
- Reduced redundant processing.
- Enhanced customer relations thro' improved knowledge.
- Enabling business process reengineering

Unit I

Data Warehousing Components

i) Data Warehouse Database

- The Central data warehouse database ② is the ~~[illegible]~~ important place of the data warehousing environment
- The database is implemented on RDBMS technology
- Data warehouse attributes such as very large database size, adhoc query processing and user creation including aggregates, multiple joins, and drill downs have become different approaches in data warehouse database

The approaches are

- Parallel relational database design that require a parallel computing platforms such as Symmetric Multiprocessors (SMPs) Massively parallel processor (MPPs) and ~~[illegible]~~ clusters of Uni(or) multiprocessor
- An innovative approach to speed up a traditional RDBMS by using new index structures to bypass relational table scans
- Multidimensional database (MDDBs) is based on the proprietary database technology. It is designed to overcome

any limitation on the data warehouse. MDDB to coupled with OLAP tools that acts as a client to the Multidimensional data store. Tools include data query, reporting, analysis and Mining tools.

(ii) Sourcing, Acquisition, Cleanup and Transformation Tools:-

* perform all of the Coversions, Summarization key changes, Structural changes and Condensation to transform data into information
* It produces programs and Control Statements including COBOL pgm, UNIX scripts, data definition Languages DDL need to move data into data warehouse from Multiple operation System.
* It also Maintains the Metadata.

The functionality includes

* Removing unwanted data from operational data Base
* Converting to Common data names and definitions
* Calculating Summaries and derived data
* Establishing defaults for missing data
* Accommodating Source definition chang

Issues

Data base heterogeneity :

* DBMS are very different on data models, data access language, data navigation, operation, integrity etc

Data heterogeneity ;

* Data is defined and used in different Models - homonyms, Synonyms, unit Compatibility, different attributes for same entity, etc

(iii) Metadata :

* Data about data
* it is used for building, Maintaining, managing and using the data warehouse
* It is of two types

(i) Technical Metadata
(ii) Business Metadata.

Technical Metadata :

* which contains information about warehouse data which is used by designers and administrators

Technical Metada documents include

- Information about data source
- Transformation description
- Warehouse object and data structure definition
- Rules used to perform data cleanup and data enhancement
- Data mapping [Source system to target database]
- Acces Authorization, backup history, archive history, information delivery history, etc.

(ii) Business Metadata.

- Contains information that gives user to understand easy way about the information stored in the data ware house.

Business Metadata includes

- Subject areas and information object type including queries, reports, images etc
- Internet home pages
- other information to support all data warehousing Components.
- Data ware house operational information eg. data history, ownership, usage of data etc

Metadata Management is provided via Metadata respositon and accompanying S/w.

* S/w is used to map source data to target database
* Component of Metadata respository is the Information directory
* The Information directory and the metadata respository ~~acts as a~~. Should be as
  * gateway to datawarehouse environment
  * easy distribution and replication of content for high performance and availability
  * Searchable by business-oriented key words
  * as a launch platform for end-user data accen and analysis tools
  * Support & Sharing of information objects such as queries, reports, data collection etc
  * Support a variety of scheduling options for request against the datawarehouse, including on-demand, one-time, event-driven etc.
  * Support and provide interface to other applications such as e-mail, Spread sheet, etc

(iv) Access Tools

It is of five Main groups

- Data Query and reporting tools
- Application development tools
- Executive Information System [EIS] Tools
- On-line Analytical processing Tools [OLAP]
- Data Mining Tools

Query and Reporting Tools

- It is divided into two groups
  - (i) reporting tools and
  - (ii) Managed query tools

Reporting tools Can be divided into

- (i) production reporting tools
- (ii) desktop report writers

Production reporting tools

- generate regular operational reports or support high-volume batch jobs such as calculating and printing pay checks

desktop report writers

- inexpensive desktop tools designed for end users

Managed Query tools

- Shield end users from the complexities of SQL and database structures by inserting a Metalayer between users and the database
- Metalayer is the S/w that provide Subject Oriented views of the database and supports point-and-click creation of SQL.

Application tools

- The application development platform integrates with OLAP tools and access all major database systems including Oracle, Sybase and Informix.
  eg. of application development environment includes, powerbuilder from powersoft, Visual basic from Microsoft

OLAP

- It is based on the concepts of Multidimensional database and allow a sophisticated user to analyze the data using Multidimensional and Complex views

to represent
rves as a
data is

* The data is organized in a Multidimensional Model which support Multidimensional data base (MDDB) & Multi Relational database (MRDB)

Data Mining.

* Process of discovering Meaningful new Correlations, patterns and trend by mining large amount of data stored on warehouse.
* It build predictive model rather than retrospective model.

Why datamining is used in organization:-

* Discover knowledge.
  * to determine hidden relationships, patterns, or correlation from data stored in on data base.
  * It can (com) used to perform. Segmentation, Classification, Association, preferencing.
* Visualize data:
  * It sense out huge amount of information stored in Corporate data base

* Correct data
    * It help to identify and correct problems in the Consistent copy

(V) Data Mart
* It means different things to different people.
* Data Store that is subsidiary to a data ware house of integrated data
* It is of two types
    * dependent data mart
    * Independent data mart

dependent data mart
* their data content is sourced from the data ware house. have high value.

Independent data Mart
* fragmented point solutions to a range business problems in the enterprise

Scanned by CamScanner

(VI) Data warehouse Administration & Mgt.

Managing data warehouse includes

- Security & priority Mgt.
- Monitoring updates from Multiple Sources
- Data quality checks
- Managing and updating meta data.
- Auditing and reporting data warehouse usage & Status
- purging data
- Replicating, subsetting & distributing data.
- Backup & recovery
- Data warehouse Storage Management.

(VII) Information Delivery System.

- used to enable the process of subscribing for data warehouse information to one or more destination by using user-specified scheduling algorithm

2. Information delivery system distributes warehouse-stored data and other information objects to other data warehouses and end user products (spreadsheets and local database.

METADATA

What is Metadata?

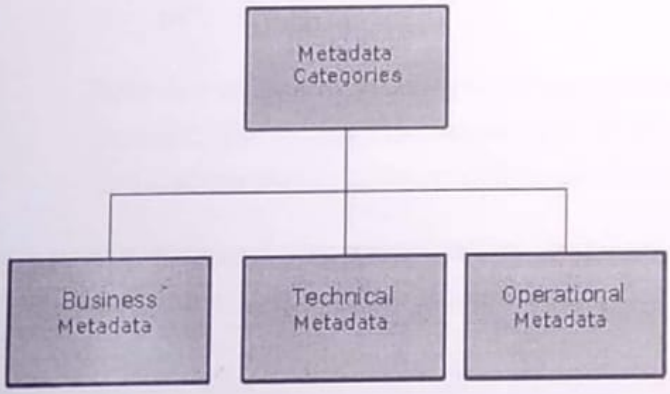
Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to detailed data. In terms of data warehouse, we can define metadata as follows.

- Metadata is the road-map to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Categories of Metadata

Metadata can be broadly categorized into three categories:

- **Business Metadata** - It has the data ownership information, business definition, and changing policies.
- **Technical Metadata** - It includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.
- **Operational Metadata** - It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.

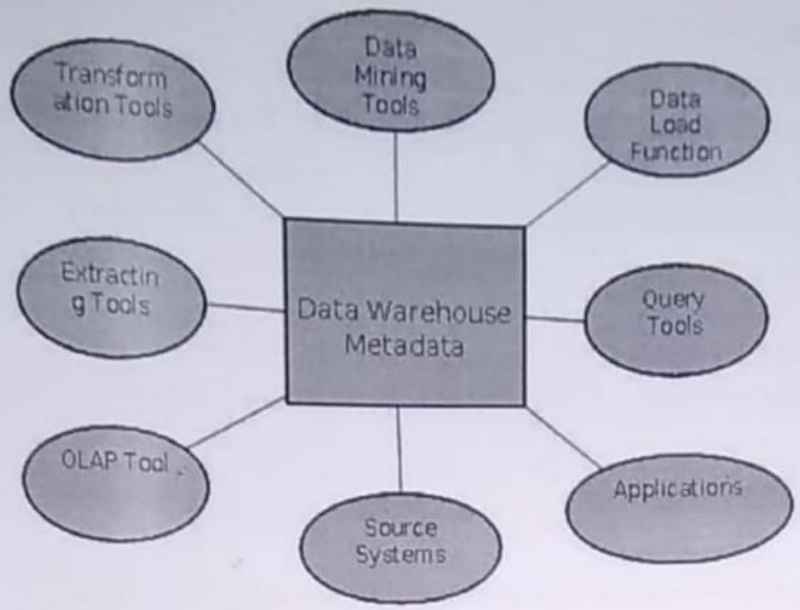


Role of Metadata

Metadata has a very important role in a data warehouse. The role of metadata in a warehouse is different from the warehouse data, yet it plays an important role. The various roles of metadata are explained below.

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.
- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata is used for query tools.
- Metadata is used in extraction and cleansing tools.
- Metadata is used in reporting tools.
- Metadata is used in transformation tools.
- Metadata plays an important role in loading functions.

The following diagram shows the roles of metadata.



Metadata Repository

Metadata repository is an integral part of a data warehouse system. It has the following metadata:

- **Definition of data warehouse** - It includes the description of structure of data warehouse. The description is defined by schema, view, hierarchies, derived data definitions, and data mart locations and contents.
- **Business metadata** - It contains has the data ownership information, business definition, and changing policies.
- **Operational Metadata** - It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.
- **Data for mapping from operational environment to data warehouse** - It includes the source databases and their contents, data extraction, data partition cleaning, transformation rules, data refresh and purging rules.
- **Algorithms for summarization** - It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

(4)

Challenges for Metadata Management

The importance of metadata can not be overstated. Metadata helps in driving the accuracy of reports, validates data transformation, and ensures the accuracy of calculations. Metadata also enforces the definition of business terms to business end-users. With all these uses of metadata, it also has its challenges. Some of the challenges are discussed below.

- Metadata in a big organization is scattered across the organization. This metadata is spread in spreadsheets, databases, and applications.
- Metadata could be present in text files or multimedia files. To use this data for information management solutions, it has to be correctly defined.
- There are no industry-wide accepted standards. Data management solution vendors have narrow focus.
- There are no easy and accepted methods of passing metadata.