UNIT -3

Building a Data Warehouse

Why a Data Warehouse Application – Business Perspectives

Several reasons :

- From a business prospective, to strive and succeed in today’s highly competitive global environment, business users demand business answers mainly because:
- Decisions need to be made quickly and correctly, using all available data
- Users are business domain experts, not computer professionals
- The amount of data increasing in the data stores, which affects response time and the sheer ability to comprehend its content.
- Competitions is heating up in the areas of business intelligence and added information value.

Why a Data Warehouse Application – Technology Perspectives

Several technology:

- First, the Data Warehouse is designed to address the incompatibility of informational and operational transactional systems.
- Secondly, the IT infrastructure is changing rapidly, and its capabilities are increasing, as evidenced by the following:
  - The prices of digital storage is rapidly dropping
  - Network bandwidth is increasing, while the price of high bandwidth is decreasing
  - The workplace is increasingly heterogeneous with respect to both the hardware and software
  - Legacy systems need to, and can, be integrated with new applications

Building a Data Warehouse

1. Business Considerations (Return on Investment)
2. Design Considerations
3. Technical Considerations
4. Implementation Considerations
5. Integrated Solutions
6. Benefits of Data Warehousing

Business Consideration

1. Approach

The Top-down Approach, meaning that an organization has developed an enterprise data model, collected enterprise-wide business requirements, and decided **to build an enterprise data warehouse with subset data marts**. (From Data warehouse to Data marts)

The Bottom-up Approach, implying that the business priorities resulted **in developing individual data marts**, which are **then integrated into enterprise data warehouse**.(From Data marts to Data Warehouse)

2. Organizational Issues

A Data Warehouse, in general, is not truly a technological issue, rather, it should be more concerned with identifying and establishing information requirements, the data sources to fulfill these requirements, and timeliness.

Design Consideration

- To be a successful, a data warehouse designer must take a holistic approach – consider all data warehouse components as parts of a single complex system and take into the account all possible data stores and all known usage requirements.

The main factors include:

- Heterogeneity of Data sources, which affects data conversion, quality, timeliness
- Use of historical data, while implies that data may be “old”.
- Tendency of databases to grow very large

Design Consideration Contd..

In addition to the general considerations, there are several specific points relevant to the data warehouse design:

- Data Content
- Metadata
- Data Distribution
- One of the biggest challenge when designing a data warehouse is the data placement and distribution strategy.
- Tools
- These tools provide facilities for defining the transformation and cleanup rules, data movement (from operational sources to the warehouses), end-user query, reporting, and data analysis.
- Performance consideration

Technical Considerations

A number of technical issues are to be considered when designing and implementing a Data Warehouse environment.

1. The Hardware Platform that would house the Data Warehouse for parallel query scalability. (Uni-Processor, Multi-processor, etc)

2. The DBMS that supports the warehouse database

3. The communication infrastructure that connects the warehouse, data marts, operational systems, and end users

4. The hardware platform and software to support the metadata repository

5. The systems management framework that enables centralized management and administration to the entire environment.

Implementation Considerations

i. Access Tools

Currently **no single tool in the market** can handle all possible data warehouse access needs. Therefore, most implementations rely on a suite of tools.

ii. Data Placement Strategies

As 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 (WORM, RAID). Second option is to **distribute data in data warehouse across** multiple servers.

Data Extraction, Cleanup, Transformation, and Migration

As a components of the Data Warehouse architecture, proper attention must be given to Data Extraction, which represents a critical success factor for a data warehouse architecture.

1. The **ability to identify data in the data source environments** that can be read by conversion tool is important.

2. **Support for the flat files**. (VSAM, ISM, IDMS) is critical, since bulk of the corporate data is still maintained in this type of data storage.

3. The **capability to merge data from multiple data stores** is required in many installations.

4. The **specification interface to indicate the data to extracted** and the conversion criteria is important.

5. The ability to **read information from data dictionaries** or import information from repository product is desired.

Contd…

iv. Metadata

- A frequently occurring problem in Data Warehouse is the problem of communicating to the end user **what information resides** in the data warehouse and how it can be accessed.
- The key to providing users and applications with a roadmap to the information stored in the warehouse is the metadata.
- It can define all data elements and their attributes, data sources and timing, and the rules that govern data use and data transformations.
- Meta data needs to be collected as the warehouse is designed and built.

Contd..

v. User Levels

- Data Warehousing is relatively new phenomenon, and a certain degree of sophistication is required on the end user's part to effectively use the warehouse. The users can be classified on the basis of their skill level in accessing the warehouse:

1. Casual Users: These **users** are most **comfortable retrieving information** from the warehouse in pre-defined formats, and running preexisting queries and reports.

2. Power Users: These users **need access tools** that combine the **simplicity of pre-defined queries and reports** with a certain degree of flexibility.

3. Experts: These **users** tend to create their own queries and perform sophisticated analysis on the information they retrieve from the warehouse.

Benefits of Data Warehouse

Successfully implemented data warehousing can realize some significance benefits which can be categorized in two categories:

1. Tangible Benefits:
   1. Product inventory **turnover** is **improved**
   2. More **cost effective decision making is enabled** by separating (ad-hoc) query processing from running against operational database.
   3. Better business intelligence is enabled by **increased quality and flexibility** of market analysis available through multi-level data structures.
2. Intangible Benefits:
   1. Improved productivity
   2. Reduced redundant processing, support, and software to support overlapping decision support applications

Critical success factor

The term "critical success factor" to refer **to those things that must go right if an undertaking is to become successful.**

For a data warehouse project, these critical success factors include,

- Set specific, achievable, and measurable goals
- Involve everyone throughout the project
- Keep an eye on the big picture
- Pay attention to the details and do not depend on assumptions
- Consider long-term strategy
- Learn from others

Requirement Analysis

- Requirement analysis is the most crucial factor for the success of any project. In the absence of a clear goal, success rates are low.

The steps in the require analysis phase:

- Clearly state the problems that have to be solved.
- Identify all data sources and the formats in which the data is stored in them.
- Identify the users of the data warehouse system.
- Clearly specify the budget in terms of time, money, and personnel.

Requirement Analysis Contd..

- Ask the users to specify their expectations from the new system.
- Ask the management to specify the success criteria.
- Filter requirements from their desires. Initially start with designing the system as per the requirements, and then later on in the enhancement phase, address the desires.
- Formulate a prioritized requirements document, listing the requirement, its source, the success criteria, and its priority.
- Get a sign-off of the requirements, resource allocation, and schedule from the top management before the team can proceed with later stages.

Planning for the data warehouse

A data warehouse is planned in terms of **business requirements, personal finances, and feasibility.**

Project Staff:

✓ Technical staff that includes the project leader, a data analyst, a business analyst, a
database administrator, and programmers who are familiar with business problems to
be solved.

✓ An ad hoc technical staff who will be called to join the project as and when needed
for specific project tasks like for technical support technical writing, training, and
helpdesk.

✓ An end-user staff that comprises subject matter experts.

✓ Corporate level sponsors such as executives from the end-user and IT community.

Project Plan

To be successful, a big project like that of a data warehouse calls for good and careful planning.

✓ An overall plan for creating the data warehouse and its infrastructure

✓ Detailed plans for every individual application that would be run

in the data warehouse environment.

Overall planning :

Overall planning :

The overall plan for creating a data warehouse includes two broad

aspects:

- Vision : Vision of the project states what must be built. Different people in the organization may have different objectives.
- Validation and estimation:

  The anticipated costs, schedule, and resources that will be

  required are estimated during the planning phase

Detailed Planning:

Moves the project from a conceptual entity to a specific one.

Aim :

- To define the budget, schedule, and intermediate and final deliverables for the data warehouse project.
- Project planning tools are used to allow managers to Visualize the time sequence in which the events must occur, the kind of personnel that will have to be assigned, and the hardware and software components that will have to be acquired and integrated.
- A well-formulated and a structured plan includes details on every step of the project, from the source of data to how the data is to be cleaned, stored, and used by the end-users, to the end-user training program.
- The training part of the plan considers teaching end-users the mechanics of how to obtain information from the warehouse, and how to go on with their need to extract strategic information from the data warehouse.

Infrastructure planning

- The infrastructure for a data warehouse includes all the hardware and software components that will be needed for the data warehouse to go live.
- The hardware components include computers, networks, terminals, or PCs; and the software components comprises of database, extraction tools, cleaning tools, and query handling.

Data warehouse design stage

The design stage of a data warehouse comprises of the following sequence:

- Design the Dimensional Model
- Develop the Architecture
- Design for Update and Expansion
- Design the Relational Database and OLAP Cubes
- Decisions in Design
- Detail Design

Decisions in Design

- Design decisions: Organization of the warehouse
- Design decisions: Back-end
- Design decisions: Data warehouse
- Design decisions: Front-end
- Design decisions: Maintaining the system

Detail Design

✓ Data warehouse capacity expansion

✓ Purging and archival of historical data.

✓ Data extraction, transformation, loading and cleansing functions

✓ Configuration management.

✓ Security.

✓ Testing of every individual module of the data warehouse and the entire system as a whole.

✓ Data refresh.

✓ Data access

✓ Data backup and recovery.

✓ Disaster recovery.

✓ Transition to production.

✓ User training and user support.

✓ Change management.

Building and implementing data marts

After completing the work of building and implementing various data marts at the department level, the next stage is to build a complete data warehouse from these data marts following a bottom-up approach.

- Test and Deploy the System
- Transition to Production
- User Training and Support
- Issues in User Support

Backup and Recovery

A data warehouse stores huge amounts of data that may have taken years to collect.

The historical data in the data warehouse may be as old as 10 or even 20 years. Before storing the data in the data warehouse, the data goes through a rigorous process of cleansing and transformation

✓ Determine what exactly has to be backed up. For this, make a list of the User tables,

system tables, and the database logs that have to be backed up.

✓ Try to make different procedures for backing up the historical and the Current data. You

may decide to back up the historical data less Frequently as compared to backing up of

the current data.

✓ Choose the appropriate medium for backing up the databases.

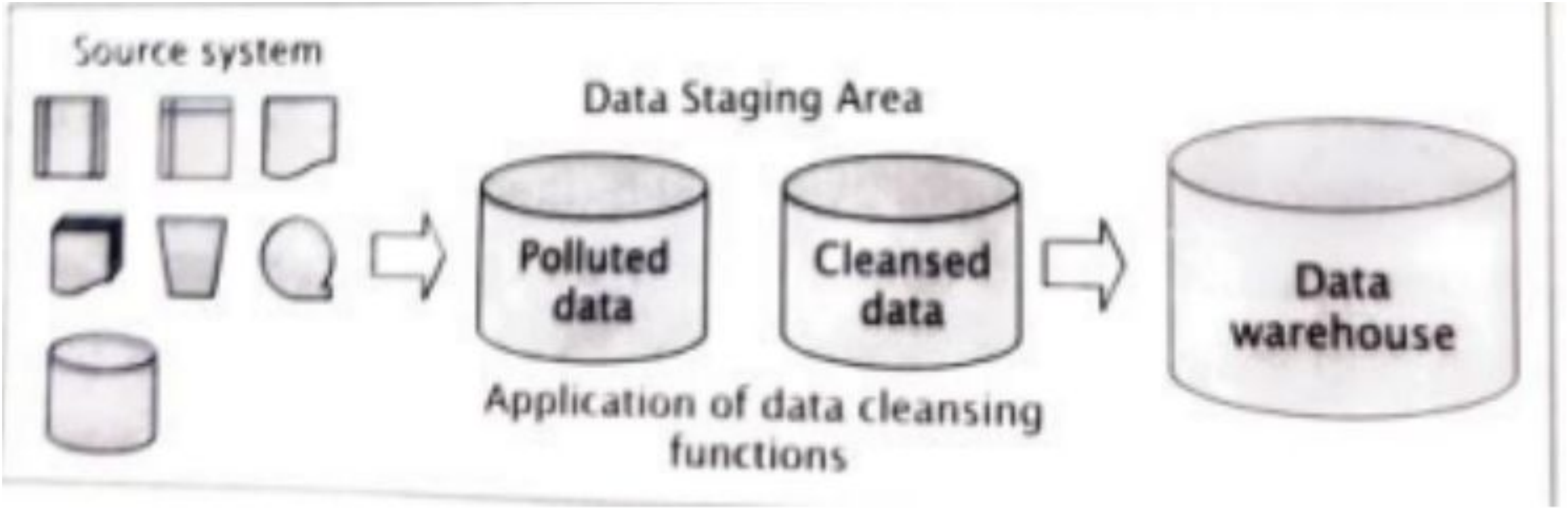
✓ Many RDBMS today make use of the container concept to hold individual files. A container is nothing but a larger storage area that can hold many files. Technically speaking, containers are also known as table Spaces, file groups, etc. RDBMS adopts special methods to efficiently back up the entire container , so you must use such features provided by RDBMS.

✓ You may choose third party tools for a high-speed backup and recovery process

✓ You must periodically archive very old data from the data warehouse. A good archival

plan helps by reducing the time for backup and restore.

Establish the data recovery quality framework



Establish the data recovery quality framework

- Data Purification To proceed with the purification process, divide the data elements into priorities with the help of users.

All the data elements into three levels of priority:

high, medium, and low.

Achieving 100% data quality is critical for the high priority data elements.

The medium-priority data requires as much cleansing as possible and some errors may be ignored to make a balance between the cost of correction and the potential effect of bad data.

The low priority data may be cleansed if you have any time and resources left.

Operating the warehouse

- Day-to-day Operations

The main usage of the data warehouse during the daytime is to service the user's queries. Other operations that can occur during the day are

- Query management.
- Backup and recovery management.
- Performance management.
- Running of housekeeping scripts.

- Warehouse Administration:

Continuous monitoring and administering activities in the data warehouse have to be done simultaneously.

Monitoring the warehouse operations

- Platform upgrades.
- Supporting end-users.
- Maintaining the metadata.
- Upgrading the warehouse.
- Addition and deletion of enhancements.
- Maintaining security.
- Managing data growth.
- ETL management.
- Storage management.
- Capacity planning
- Information delivery enhancements.

Data warehouse pitfalls

- The following list summarizes the Much time limitations of a warehouse.
- Much time is wasted in data extracting, cleaning, and loading of data.
- One thing that should always be taken as default is that the scope of the data warehouse will always go beyond expectations.
- There will be a number of problems that the team will have to face because of the **disparate source systems** that feed the data in the data warehouse
- Often there will be a need to store data not being captured by the existing systems.
- In continuation with the previous problem, there would also be validate the data that is currently not being validated by the transaction processing systems.

Data warehouse pitfalls

- A data warehouse is a high-maintenance system. The data warehouse project team will fail if it concentrates more on resource and optimization neglect, data and customer management issues and an understanding of what adds value to the customer.

Meta Data – Introduction

- Metadata defines the contents and location of the data (or data model) in the data warehouse, relationships between the operational database and the data warehouse and the business views of the data in the warehouse as **accessible to the end-user tools**
- it acts as logical link between the decision support system application and the data warehouse

Types of Metadata

- Operational Metadata
- Extraction and Transformation Metadata
- End-User Metadata