

BUILDING A DATA WAREHOUSE

WHY DO WE NEED TO BUILD A DATA WAREHOUSE?

- To store lots of historical data in one place
- Quicker access
- Easier to integrate many different sources of data in applications

New business and technology drivers often make building a data warehouse a strategic imperative.

BUSINESS CONSIDERATIONS: RETURN ON INVESTMENT

APPROACH:

- TOP DOWN APPROACH: An organisation has developed an enterprise data model, collected business requirements and decided to build a data warehouse with subset data marts.
- BOTTOM UP APPROACH: The business priorities resulted in developing individual data marts are then integrated into a data warehouse.

ORGANISATIONAL ISSUES:

- The organisation needs to bring together data that cuts across a company's operational systems as well as data from outside the company
- It should be concerned with identifying and establishing information requirements and timeliness.

DESIGN CONSIDERATIONS

- Heterogeneity of data sources
- Use of historical data
- Tendency of databases to grow very large

Data warehouse design must focus on being business driven and is never finished.

DATA CONTENT:

- Data model is the template that describes how information will be organised within the integrated warehouse framework.
- It identifies relationships including keys, attributes and groupings.
- The data mart is cleaned up, transformed and consistent with the data warehouse and other data marts from the same warehouse.

METADATA:

- Defines contents and location of data in the warehouse.
- Provides a decision-support-oriented pointer to warehouse data.

DATA DISTRIBUTION

- Data placement and distribution design should consider several options such as subject area, location.
- Distribution also creates problems such as data buildup on the network.

TOOLS:

- Provide facilities for defining the transformation and cleanup rules, data movement and data analysis.
- Have to make sure that all tools are compatible with the warehouse

PERFORMANCE CONSIDERATIONS:

- Should support interactive query processing
- Difficult to predict efficiency of querying due to unpredictable usage patterns
- Data warehouses have to be built for a particular need.

NINE DECISIONS IN THE DESIGN OF A DATA WAREHOUSE:

- Difficult to create data warehouse when there are so many questions to answer.
- To help this process there are some generalised decisions to take

1. Choosing the subject matter
2. Deciding exactly what a fact table record represents
3. Identifying and conforming the dimensions
 - i. A well architected set of dimensions makes the data mart understandable and easy to use.
 - ii. For 2 data marts to share information they should be of the same dimensions
4. Choosing the facts
5. Storing precalculations in the fact table
6. Rounding out the dimension tables
7. Choosing the duration of the database
8. The need to track slowly changing dimensions
9. Deciding the query priorities and the query modes.

TECHNICAL CONSIDERATIONS:

- The hardware platform that would house the data wardrobe.
- The DBMS that supports the warehouse database
- The communication infrastructure that connects the warehouse, data marts, operational systems and end users
- The hardware platform and the software to support the metadata repository
- The systems management framework that enables the centralized management and administration of the entire environment.

HARDWARE PLATFORMS:

- Decide which platform is best to build a warehouse
- Should have capacity for handling the data
- Has to be specialized for all tasks associated with the data warehouse
- Balanced approach: All components must be balanced and thus the resulting system forms a scalable platform
- Optimal hardware for parallel query scalability:

DATA WAREHOUSE AND DBMS SPECIALIZATION

- There are some databases that are optimized specifically for data warehousing
- Eg: Red Brick Warehouse

COMMUNICATIONS INFRASTRUCTURE:

- A data warehouse user requires a large bandwidth to interact with the data warehouse.
- Communications networks have to be expanded and new hardware and software has to be purchased.

IMPLEMENTATION CONSIDERATIONS:

- Collect and analyse business requirements
- Create a data model and a physical design for the warehouse
- Define data sources
- Choose the database technology and platform
- Extract the data from the operational database, clean it and load it into the database.
- Choose database access and reporting tools
- Choose database connectivity software
- Choose data analysis and presentation software
- Update the data warehouse

ACCESS TOOLS:

- Simple tabular form reporting
- Ranking
- Multivariable analysis
- Statistical analysis
- Complex queries with multitable joins and sophisticated search

DATA EXTRACTION, CLEANUP, TRANSFORMATION AND MIGRATION

- Have to identify data that can be read by the conversion tool
- Support for flat files, index files is critical
- Capability to merge data from multiple data stores
- Ability to read data from data dictionaries
- The capability to create summarization, aggregation and derivation records

VENDOR SOLUTIONS:

Some vendors are more focussed on fulfilling requirements of a warehouse

PRISM SOLUTIONS:

Maps data to a target DBMS to be used as a warehouse

CARLETON'S PASSPORT:

Product that consists of 2 components. One collects the file record table and converts them to the Passport Data Language. The second creates the extracts.

INFORMATION BUILDERS INC. :

Provides relational view and a uniform SQL access to relational and non relational data.:

SAS INSITUTE INC:

Uses its tools as a means to serve all data warehousing functions

DATA PLACEMENT STRATEGIES

- Store the data in some data storage
- Distribute the data in the warehouse across multiple servers.

DATA REPLICATION:

- Store the same data in multiple locations
- Store frequently used data separately

DATABASE GATEWAYS:

- Provides users with the ability to access small amounts of mainframe data.

METADATA:

- Needs to be collected as the warehouse is built
- Same metadata needs to be available to all tools
- A well thought out strategy is required.

USER SOPHISTICATION LEVELS:

- CASUAL USERS: Most comfortable retrieving information in predefined formats
- POWER USERS: Combine predefined queries with ad hoc queries that they created
- EXPERTS: Create their own complex queries

INTEGRATED SOLUTIONS:

A number of vendors participate in data warehousing by providing a suite of services and products that go beyond one particular component of the data warehouse

Eg: Digital Equipment Corp., Hewlett Packard, IBM, Sequent

BENEFITS OF DATA WAREHOUSING:

TANGIBLE BENEFITS:

- Product inventory turnover is improved
- Costs of product introduction is decreased.
- More cost effective decision making is enabled
- Better business intelligence is enabled by increased quality and flexibility of market analysis
- Enhanced asset and liability management

INTANGIBLE BENEFITS:

- Improved productivity by keeping all data in a single location.
- Reduced redundant processing, support and software to support overlapping decision support applications
- Enhanced customer relations through improved knowledge of customer trends
- Enabling business process reengineering