

D2A - Planning & Requirements, Infrastructure

Planning

A Data Warehouse Project

Expectation

- Is the intended data warehouse right for the organization?
- What does the organization get from the data warehouse?

Risk

Need risk analysis on benefit, losses, opportunity

Development Style

- Bottom-up
- Top-down

Build or Buy?

Build by self or buy from outside.

One or More Vendor?

SAS?SQL?MicroSoft?

There are Advantages and disadvantages for both

Business Requirements

These requirements (strategic information needs) drive the data warehouse

Top-Level Management Support

How much is the support from top-level management?

Cost of a DW Project

Roughly 31% hardware, 24% software, 35% staff, 10% administration

Justify the DW

- Its tangible and intangible
- Real benefits may not be seen until the entire DW is build and is in fully use.
- Many organizations do a DW project without full cost-justification
- **3 Example Method:**
 - Compare current cost of supporting decision-making with the predicted cost of building a DW to support decision-making.
 - Calculate the predicted business value of the DW, such as increase in sales, increase in profit, increase in market share...,to see is it bigger than the cost to build a DW.
 - Identify the organization components that affect the DW or are affected by the DW. Calculate the costs, like hardware, vendor software, in-house software, supporting staff, maintenance...Then calculate the benefits, including tangible and intangible benefits, like cost reduction, profit increase, business effectiveness increase, put a dollar value on each benefit. Compare the cost and benefit.

DW Project Tech Design

Data Acquisition

- From internal or external sources
- Need transformation, cleansing
- Data quality issue

Data Storage

- Large volume of data
- Need extra staging space
- May rapidly grow
- Structured and unstructured data
- Can use parallel processing

Information Delivery

- Lots of users
- Different users have different requirements
- Query may be complex
- Should support multidimensional analysis
- DW can be web-based
- Should manage metadata
- Multiple vendor may access data at same time

DW Project Danger Sign

- The requirements definition step is well past the target date
- Need to write too many in-house programs (means have to look at other tools)
- Users do not cooperate to provide details of data (need top-level support)
- Users are not comfortable with query tools (ensure tool is right for jobs, then more training)
- Data quality problem bring over to the ETL process

A Good DW Project

- Support from top-level management
- Manage users' expectation
- Get users enthusiastically involved
- Do not under-estimate the time and effort for ETL
- Selection of architecture first, then technology, then tools
- Use a easy-to-use query tool
- Plan for growth
- Assign a user-oriented project manager
- Focus the design on query
- Only load the data that is needed

Defining Requirements for DW

Talk to

- **Senior executives:** organizational objective, success criteria
- **Key department manager:** department objective, success metrics, business dimension
- **Business analyst:** business issue, analysis dimension
- **Operational system DBA:** source data, information delivery, data quality, routine analysis

Requirements Driving the DW Development

- Business requirements of users are the most powerful driving force for the development of a DW
- Business requirements guide, influence, direct many decisions about DW project.
- Because: DW is an information delivery system, users access and create un-predefined output, need right information in optimal format.

Infrastructure

We need infrastructure to support our architecture.

Operational Infrastructure 操作型设施

- Software tools for ETL, DW
- DW project team + Administration staff
- Business rules

Physical Infrastructure 物理设施

- Use as much of existing physical infrastructure as possible
- Keep the physical infrastructure as modular as possible
- When required and when the newer hardware is cheap enough, replace the old one
- Need to think about how much disk space, how much processors...

Physical Infrastructure Components

- Desktop clients
- Application servers to ensure the running of tools (control, manage metadata, network connection, OLAP)
- Data staging machine
- Development machine
- Data mart storage
- Data warehouse storage

Technical Consideration

- Hardware and OS
- How to handle loads (eg. clustering, symmetric parallel processing..)
- Query governor (abort runaway query)
- Query optimizer (parse and optimize query)
- Query management (balance the execution of different type of query)
- Load utility (high-performance data loading)
- Metadata management
- Scalability (larger number of users and data volume)

- Extensibility
- Portability (across platform)
- Query toll Application Program Interface (API)
- Administration

Data Movement

- Data movement is very dependent on infrastructure design

Options

1. Shared disk
2. Mass transmission
3. Realtime connection
4. Manual methods

10 Tool Type

1. Data Modelling Tool

- Help to create and maintain data models of source database and DW
- Able to build dimensional models in star schema

2. Data Extraction Tool

- Support bulk data extraction for full refreshes
- Support revision data extraction for incremental loading
- Related to platforms of source systems

3. Data Transformation

- Can convert the format and type of extracted data
- Can provide specified default value
- Can split, combine, standardize, de-duplicate data.

4. Data Loading

- Support full refreshes
- Support incremental loading
- Can create PK for DW tables

5. Data Quality

- Help to find and correct errors
- Can be used in source systems or staging areas.
- Can generate data quality reports based on some measurements

6. Query and Reporting

- Support complex query
- Help users to write query and generate report
- Optimize query for better execution
- 分report writer and report server

7. OLAP

- Support complex query
- Provide pre-defined query

8. Alert System

- Highlights pre-defined exceptions
- Provide alerts from DW data to support decision-making
- Three basic alert types are: from individual source systems, from integrated enterprise-wide data warehouses, and from individual data marts.

9. Middleware and Connectivity

- Transparent access to source systems
- Transparent access to databases which have different types and are on multiple platforms

10. DW Management

- Support daily maintain needs
- Able to track loading processes
- Able to track query response time to improve performance