BUILDING A DATA WAREHOUSE

V. Vaishnavee

AP – AI-DS

SNSCE

Business Considerations: return on investment

Approach

Organizations interested in development of a data warehouse can choose one of the following two approaches:

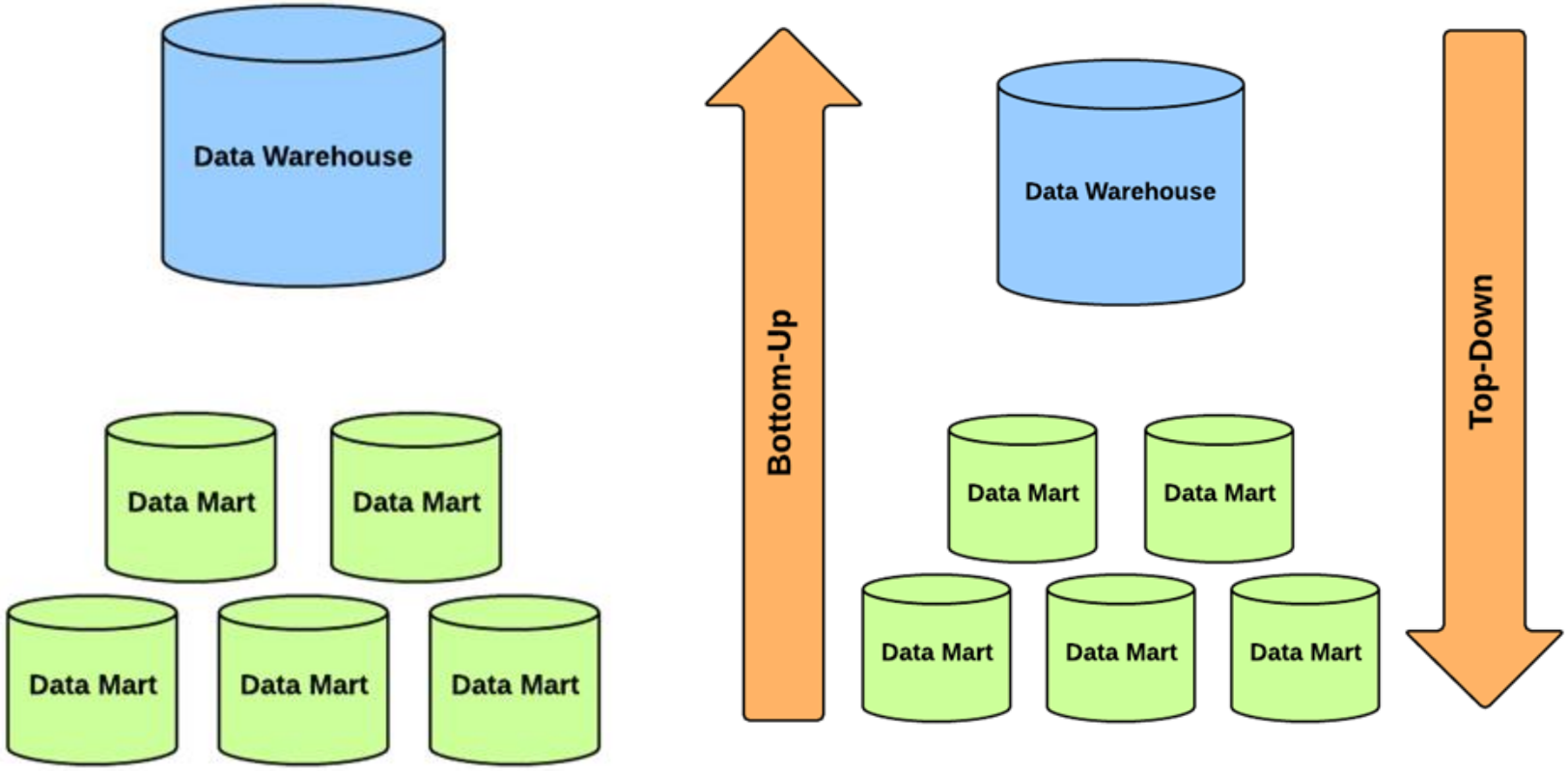
- Top - Down Approach (Suggested by Bill Inmon)
- Bottom - Up Approach (Suggested by Ralph Kimball)

Top-Down Approach:

Organization has developed enterprise data model, collected enterprise wide business requirements, and decided to build an enterprise data warehouse (EDW) with subset data marts.

Bottom –up approach:

Business priorities resulted in developing individual data marts, which are then integrated into enterprise data warehouse (EDW).


Data Warehouse
Data Mart
Data Mart
Data Mart
Data Mart
Data Mart
Bottom-Up
Data Warehouse
Top-Down
Data Mart
Data Mart
Data Mart
Data Mart
Data Mart

Design Considerations

- In addition to the general considerations there are following specific points relevant to the data warehouse design:

1. Data content

- The content and structure of the data warehouse are reflected in its data model.
- The data model is the template that describes how information will be organized within the integrated warehouse framework.
- The data warehouse data must be a detailed data. It must be formatted, cleaned up and transformed to fit the warehouse data model.

2. Meta data

- It defines the location and contents of data in the warehouse.
- Meta data is searchable by users to find data definitions or subject areas.
- In other words, it must provide decision support oriented pointers to warehouse data and thus provides a logical link between warehouse data and decision support applications.

Design Considerations

3. Data distribution

- One of the biggest challenges when designing a data warehouse is the data placement and distribution strategy.
- Data volumes continue to grow in nature.
- Therefore, it becomes necessary to know how the data should be divided across multiple servers and which users should get access to which types of data.
- The data can be distributed based on the subject area, location (geographical region), or time (current, month, year).

4. Tools

- A number of tools are available that are specifically designed to help in the implementation of the data warehouse.
- All selected tools must be compatible with the given data warehouse environment and with each other.
- All tools must be able to use a common Meta data repository.

Design Considerations - **Design steps**

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

1.Choosing the subject matter

Function refers to subject matter of a particular data mart, Ex: Bill Payment Process

2. Deciding what a fact table represents

decide what a record of the fact table is to represent i.e the grain. Ex: grain is the single payment

3. Identifying and conforming the dimensions

Dimensions set the context for asking questions about the facts in the fact table. Ex: who made the bill payment

4. Choosing the facts

Facts should be numeric and additive

5. Storing pre calculations in the fact table

once facts have been selected each should be re-examined to determine whether there are opportunities to use pre-calculations

6. Rounding out the dimension table

what properties to include in dimension table to best describe it.

7. Choosing the duration of the db

how long to keep the data

8. The need to track slowly changing dimensions

where a changed dimension attribute is overwritten.

9. Deciding the query priorities and query models

indexing for performance, indexed views, partitioning, physical sort order.

Storage, backup, security

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
  - The communication infrastructure that connects data marts, operational systems and end users
  - The hardware and software to support meta data repository
  - The systems management framework that enables admin of the entire environment

Implementation considerations

- The following logical steps needed to implement a data warehouse:
  - Collect and analyze business requirements
  - Create a data model and a physical design
  - Define data sources
  - Choose the DB tech and platform
  - Extract the data from operational DB, transform it, clean it up and load it into the warehouse
  - Choose DB access and reporting tools
  - Choose DB connectivity software
  - Choose data analysis and presentation s/w
  - Update the data warehouse

Implementation considerations

1. Access tools

- Data warehouse implementation relies on selecting suitable data access tools.
- The best way to choose this is based on the type of data can be selected using this tool and the kind of access it permits for a particular user.
- The following lists the various types of data that can be accessed:
    - Simple tabular form data
    - Ranking data
    - Multivariable data
    - Time series data
    - Graphing, charting data
    - Complex textual search data
    - Statistical analysis data
    - Data for testing of hypothesis, trends and patterns
    - Predefined repeatable queries
    - Reporting and analysis data Complex queries with multiple joins, multi level sub queries and sophisticated search criteria

Implementation considerations

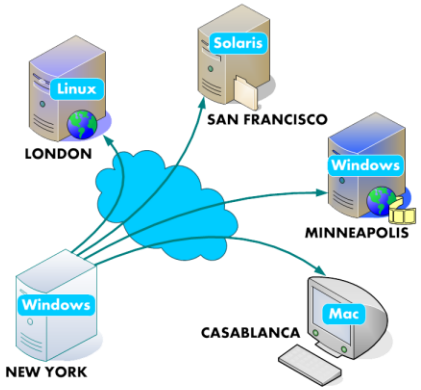
2. Data extraction, clean up, transformation and migration

- 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

Implementation considerations

3. 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.

Implementation considerations

4. 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 information
  - Presentation of information
  - Testing of hypothesis
  - Discovery of information
  - Sharing the analysis

Benefits of data warehousing

- The benefits can be classified into two:
- **Tangible benefits (quantified / measureable):It includes,**
  - Improvement in product inventory
  - Decrement in production cost
  - Improvement in selection of target markets
  - Enhancement in asset and liability management
- **Intangible benefits (not easy to quantified): It includes,**
  - Improvement in productivity by keeping all data in single location and eliminating rekeying of data
  - Reduced redundant processing
  - Enhanced customer relation


THE END