DATA WAREHOUSING

Ofori Boateng, PhD
Professor, University of Northern Virginia
BMGT531 – 1900- SU 2011 Business Intelligence Project

Jagir Singh, Greeshma, P Singh
University of Northern Virginia

Abstract

Data Warehousing is a technique of bringing collectively all of a company’s data from different computer systems, together with those connecting to customers, employees, vendors, product, inventory, and financial. The data warehouse connects different database together in order to offer a more inclusive data set for making decision. The paper considers how different ways of determining such warehouse have been developed and how confident organizations have used them to increase control over data and decision making. This reveals that organization that can develop a strong system, data warehousing is value the cost. A physical repository where relational data are specially organized to provide enterprise, cleansed data in a standardized format.

Data Warehousing Process Overview

- Organization constantly collects data, information and knowledge at progressively more accelerated rate and stores them in computerized system.
- The number of users desire to access the information continues to increase as a result of improved reliability and availability of net.

![Data Warehouse Framework and Views](image-url)
Operationnal Responsibility of Data Warehousing

The majority of the data warehousing operation report to traditional information technology units of a variety of titles. Other unit titles liable for data mining use were: Data Administration, Data warehousing team, Office of institutional Research and planning, Administrative Applications, and a statistical Consulting Center. Data Warehousing were programmed by Twenty one respondents as under the purview of information system. One institution is a part of a multi university data resource system. Several institutions have very particular units such as Data warehouse/ Data administration and Data warehouse team.

Data Warehouse Processing

![Architecture of Web-Based Data Warehousing](image1)

![Architecture of a Three-Tier Data Warehouse](image2)
Data Warehousing Architectures

Following are the factors that affect the architectures selections decision:
- Nature of end-user tasks
- Information interdependence between organizational units
- Social/political factors
- Constraints on possessions
- Professed ability of the in-house IT staff
- Upper management’s information needs
- Necessity of need for a data warehouse
- Strategic view of the data warehouse former to implementation
- Compatibility with existing system

Following are the issues to consider when deciding which architecture to use:
- What tools will be used to sustain data recovery and analysis?
- Which database management system should be used?
- Will data migration tools be used to load the data warehouse?
- Will parallel processing or partitioning be used?

Data Mining Capabilities

Regardless of how smartly and productively the information management system is planned, built and operated; the information management system is basically a repository, or a storage facility. The value is completely dependent on the analytic applications that access, process, and present the data, information and knowledge to sustain research and problem solving necessities. This is the process of data mining.

Data Warehouse Vendors

Following are the six guidelines to be considered while making a vendor list:
- Industry experience
- Market Share
- Financial strength
- Qualified consultants
- ERP linkages
- Established partnerships

Data Integration and the Extraction

Data Integration

Data Integration comprises of the three major processes:
- Data access
- Data federation
- Change capture
When these three processes are correctly implemented, data can be accessed and made available to an array of ETL and analysis tools and data warehousing environments.

**Enterprise Application Integration**

It can define as the technology that provides a medium for pushing data from source system into a data warehouse.

**Enterprise Information Integration**

It can be define as an evolving tools space that promises real time data integration from number of sources, such as relational databases, multidimensional database, and web services.

**Extraction, Transformation and Load**

It can be define as the data warehousing process that consists of extraction, transformation and load which is called putting data into the data warehouse.

Following are the major points in selecting an ETL tool:
- Capability to read from and write to an infinite number of data source architectures.
- Automatic capturing and delivery of metadata.
- A history of meeting the requirements to open standards.

**Data Warehouse Development**

Following are the major points of direct benefit of data warehouse:
- Simplification of data access
- Better and timelier information
- Allows end users to perform extensive analysis
- Enhanced system performance.
- Allows a consolidated view of corporate data

Following are the major points of indirect benefit resulting from end users using these direct benefits:
- Present competitive advantage
- Enhance customer service and satisfaction
- Facilitate decision making
- Help in reforming business process

Data warehouse development approaches:
- Inman Model: EDW approach
- Kimball Model: data mart approach

Some best practices for implementing a data warehouse:
- Project must fit with corporate strategy and business objectives.
- There must be complete buy-in to the project by executives, managers, and users.
- It is important to manage user expectations about the completed project.
- The data warehouse must be built incrementally.
- Build in adaptability.
Some best practices for implementing a data warehouse (Weir, 2002):
• The project must be managed by both IT and business professionals.
• Develop a business/supplier relationship.
• Only load data that have been cleansed and are of a quality understood by the organization.
• Do not overlook training requirements.
• Be politically aware.

Failure factors in data warehouse projects:
• Cultural issues being ignored.

- Inappropriate architecture.
- Unclear business objectives.
- Missing information.
- Unrealistic expectations.
- Low levels of data summarization.
- Low data quality.

**Grain:** It may be defined as the highest of detail that is supported in a data warehouse.

**Drill-down:** It can be defined as the process of inquiring beyond a summarized value to examine each of the detail transaction that comprises the summary.

Following are the Implementation factors that can be categorized into three criteria:
• User participation in the development of data and access modeling is a critical success.
• Factor in data warehouse development.
• Organizational issues.
• Project issues.
• Technical issues.

**Real Time Data Warehousing**

**Real-Time Data Warehousing**

It can be defined as the process of loading and providing data via a data warehouse as they become available.

**Levels of Data Warehouses**

• Reports what happened.
• Some analysis occurs.
• Provides prediction capabilities.
• Operation alization.
• Becomes capable of making events happen.
The need for real-time data
- A business often cannot afford to wait a whole day for its operational data to load into the data warehouse for analysis.
- Provides incremental real-time data showing every state change and almost analogous patterns over time.

- Maintaining metadata in sync is possible.
- Less costly to develop, maintain, and secure one huge data warehouse so that data are centralized for BI/BA tools.
- An EAI with real-time data collection can reduce or eliminate the nightly batch processes.
Data Warehouses in Opposition to Operational System

Operational system are optimized for preservation of data integrity and speed of recording ensure database designs often result in information from a business transaction being stored in dozen to hundreds of tables. Relational database are efficient at managing are relationship between these tables.

Data warehouse are optimized for speed of data analysis. Frequently data in warehouses are demoralized via a dimension based model. Also, to speed data retrieval, data warehouse data are often stored multiple times- in their most granular form and in summarized forms called aggregates.

Benefit of Data Warehousing

- A data warehouse provides a common data model for all data of interest regardless of the data’s source. This makes it easier to report and analyze information than it would be if multiple data models were used to retrieve information such as sales invoices, order receipts, general ledger charge.
- Prior to loading data into the data warehouse, inconsistencies are identified and resolved. This greatly reporting and analysis.
- Information in the data warehouse is under the control of data warehouse users so that. Even if the sources system data are purged over time.
- Because they are separate from operational systems, data warehouse provide retrieval of data without slowing down operational system.
- Data warehouse can work in conjunctions with and, hence, enhance the value of operational business application.

Disadvantages of Data Warehousing

- Data warehouses are not the optimal environment for unstructured data.
- Because data must be extracted. Transformed and loaded into the warehouse, there is an element of latency in data warehouse data.
- Over their life, data warehouse can have high costs.
- Data warehouse can get outdated relatively quickly.
- There is often a fine line between data warehouses and operational system. Duplicate expensive functionally may be developed.

Massive data warehouses and scalability
The Main Issues Pertaining to Scalability

- The amount of data in the warehouse.
- How quickly the warehouse is expected to grow.
- The number of concurrent users.
- The complexity of user queries.
- Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse.

Data Warehouse Administration

Data Warehouse Administrator (DWA)

A person responsible for the administration and management of a data warehouse

Effective security in a data warehouse should focus on four main areas:

- Establishing effective corporate and security policies and procedures.
- Implementing logical security procedures and techniques to restrict access.
- Limiting physical access to the data center environment.
- Establishing an effective internal control review process with an emphasis on security and privacy.

Following are the issues to consider building a successful data warehouse:

- Delivering data with overlapping and confusing definitions.
- Believing promises of performance, capacity, and scalability.
- Believing that your problems are over when the data warehouse is up and running.
- Focusing on ad hoc data mining and periodic reporting instead of alerts.
Data Warehouse Model

There are many different models of data warehouses. Online Transaction Processing, which is a data warehouse model, is built for speed and ease of use. Another type of data warehouse model is called Online Analytical processing, which is more difficult to use and adds an extra step of analysis within the data. Usually it requires more steps which slows the process down and requires much more data in order to analyze certain queries.

In addition to this model, one of the more common data warehouse models include a data warehouse that is subject oriented, time variant, non volatile and integrated. Subject oriented means that data is linked together and is organized by relationships. Time variant means that any data that is changed in the data warehouse can be tracked. Usually all changes to data are stamped with a time-date and with a before and after value, so that you can show the changes throughout a period of time. Non volatile means that the data is never deleted or erased. This is a great way to protect your most crucial data. Because this data is retained, you can continue to use it in a later analysis. Finally, the data is integrated, which means that a data warehouse uses data that is organizational wide instead of from just one department.

Besides the term data warehouse, a term that is frequently used is a data mart. Data marts are smaller and less integrated data housings. They might be just a database on human resources records or sales data on just one division

- Characteristics of data warehousing.
- Subject oriented
- Integrated

Types of Data Warehouse

- With improvements in technology, as well as innovations in using data warehousing techniques, data warehouses have changed from Offline Operational Databases to include an Online Integrated data warehouse.
- Offline Operational Data Warehouses are data warehouses where data is usually copied and pasted from real time data networks into an offline system where it can be used. It is usually the simplest and less technical type of data warehouse.
- Offline Data Warehouses are data warehouses that are updated frequently, daily, weekly or monthly and that data is then stored in an integrated structure, where others can access it and perform reporting.
- Real Time Data Warehouses are data warehouses where it is updated each moment with the influx of new data. For instance, a Real Time Data Warehouse might incorporate data from a Point of Sales system and is updated with each sale that is made.
- Integrated Data Warehouses are data warehouses that can be used for other systems to access them for operational systems. Some Integrated Data Warehouses are used by other data warehouses, allowing them to access them to process reports, as well as look up current data.
**Data Warehouse & Data Mart**

A data warehouse is a relational database that is designed for query and analysis rather than transaction processing. A data warehouse usually contains historical data that is derived from transaction data. It separates analysis workload from transaction workload and enables a business to consolidate data from several sources.

In addition to a relational database, a data warehouse environment often consists of an ETL solution, an OLAP engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

**There are Three Types of Data Warehouses**

- **Enterprise Data Warehouse** – An enterprise data warehouse provides a central database for decision support throughout the enterprise.
- **ODS (Operational Data Store)** – This has a broad enterprise-wide scope, but unlike the real enterprise data warehouse, data is refreshed in near real time and used for routine business activity. One of the typical applications of the ODS (Operational Data Store) is to hold the recent data before migration to the Data Warehouse. Typically, the ODS are not conceptually equivalent to the Data Warehouse albeit do store the data that have a deeper level of the history than that of the OLTP data.
- **Data Mart** – Data mart is a subset of data warehouse and it supports a particular region, business unit or business function.
Conclusion

In this paper we have described a met model for data warehouse operational processes and techniques to design, administrate and facilitate the evolution of the data warehouse through the exploitation of the entities of this met model. This met model takes advantage of the clustering of its entities in logical, physical and conceptual perspectives, involving a high level conceptual description, which can be linked to the actual structural and physical aspects of the data warehouse architecture. This approach is integrated with the results of previous research, where data Warehouse architecture and quality met models have been proposed assuming the same categorization.

The physical perspective of the proposed met model covers the execution details of Data warehouse processes. At the same time, the logical perspective is capable of modeling the structure of complex.

References
