RESEARCH ISSUES ON DATA WAREHOUSE MAINTENANCE

[1]S. Sai Satyanarayana Reddy, [2]A.Lavanya, [3]Dr.V.Khanna,[4] Dr.L.S.S.Reddy

Research Scholar Student Supervisor & Guide internal Guide

[1] PadmaSri Dr. B.V.Raju Institute of Technology, Vishnu Pur, Narsapur, Medhak (DT), A.P, India
Email saisn90@gmail.com, , phani_lav@yahoo.com meetss90@yahoo.in Phone:-91-9440012540

Abstract

Abstract: *Data Warehousing has been a buzz word in the industry. Researches have been constantly involved in finding new ways of designing and developing Data Warehousing Architectures, algorithms and tools for bringing together some selected data from multiple heterogeneous databases into a single repository. The real work of taking output from the data warehouse depends largely on how it is managed. Although a lot of research is going on to enhance the design and development of Data warehouse, Very little effort has been spent on the maintenance side. Without proper maintenance data warehouse is not going to give the desired output which is expected of it. In this work a comparison of various architectures of Data Warehousing system, based on analyzed on concepts like wrapper, monitor, integrator, metadata, data quantity indicator is considered.*

The Objective of the work is to come out with comparisons between theoretical projections and real world findings on all these factors. The Final goal of the work is to suggest an improved methodology for potential users of the data warehouse in their decision making process in the Business process system.

Keywords: data warehousing, maintenance, warehousing problems, etc

INTRODUCTION

In the current scenario of changing business conditions organization's management needs to have access to more and better information. Most organizations are now days operating using information technology as the backbone of their operations but the fact is that despite having a large number of powerful desktop and notebook computers and a fast and reliable network, access to information that is already available within the organization is very difficult or otherwise not possible . All organizations whether large or small using Information Technology for the operations produce large amount of data about their business including data about sales, customers, products, services and people. But in most cases this data remains in the operational systems and can't be used by the organization. This phenomenon is called 'data in jail'. Experts say that only a small portion of this data that is entered, processed and stored is actually available to decision makers and management of the enterprise. The unavailability of this data can cause significant reduction in sales and profits of organizations and vice versa.

In the latter half of the 20th century, there existed a large number and types of databases .Many large businesses found themselves with data scattered across multiple platforms and variations of technology, making it almost impossible for any one individual to use data from multiple sources.

.

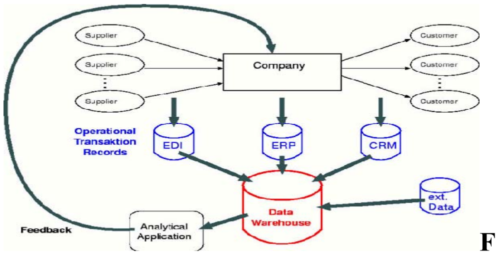


Fig 1.1
Data Warehouse in an Organization

1.2 Data Warehouse Market

The concept of data warehousing was in the industry since the early 1980's but during the early 90's it's real importance was recognized. Since than virtually every global 2000 company has acquired some form of data warehousing technology y and is using it in some form for decision support .
Each vendor that joins the battle is hoping to address the concerns of at least a slice of what is estimated to be a $4Billion market currently, and which will grow to an estimated $9.9Billion b y 2008. Through 2006, Meta Group expects to see vendors increase their sales, marketing, and development focus on this market as the transaction processing market recedes in emphasis. This will mean larger services organizations for some vendors, while others will consider solidifying or expanding relationships with third-party value-added

978-0-7695-3516-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICACC.2009.144

resellers. Some consolidation of end-user business intelligence tools (e.g., Business Objects, Cognos), extract/transform/load vendors (e.g. Informatica, Ab Initio), or boutique services firms that specialize in data warehousing is also likely to be seen through 2008.

1.3 Data Warehouse

According to a data warehouse is a subject oriented (high level entities of enterprise for e.g. customer, product), integrated (consistent naming convention, consistent variables, consistent attributes of data), time variant (data obtained over a long period of time), and nonvolatile (arrival of new data doesn't updates previous data) collection of data to support the management's decision making. The data that enters the data warehouse comes from the traditional operational systems working in the enterprise ibid.

A data warehouse is a copy of transaction data specifically structured for querying and reporting . It is a huge (sometimes terabytes of disk storage) database, which stores volumes of historical data for the company. The concept of a data warehouse came into existence as a result of two different sets of requirements . First, the end users need to view and understand the company wide view of information and second, the information system (IS) department's need to manage the data for technological and economic reasons.

1.5 Data Warehouse Maintenance Problem

As data warehousing is an emerging area, a lot of problems are found in the system. One of the major problems faced by the industry today is data warehouse maintenance. As data warehouses are huge systems. Lot of time is spent on data extraction, cleansing and loading process. Experts say usually 80% of the time building a data warehouse is consumed b y these tasks. As the users of the data warehouse experience the capabilities of the data warehouse their demands will increase gradually. The developers of data warehouse often find problems in the operational systems from where data must be captured.

1.6 Problems after Deployment of Data Warehouse

Some of the common problems faced after the deployment of a data warehouse include

1. Some times after the deployment of data warehouse it may be needed to delete some useless data. Someone has to make a decision which data to delete and which one to keep. The usual cause for this problem is the storage cost.
2. In a data warehouse queries to retrieve information from data warehouse need to be written. Someone has to decide which queries should be user written and which should be written by the information system.
3. After the deployment of a data warehouse, the users will find a lot of loop holes where there are opportunities to fine tune the data warehouse.
4. The users of the data warehouse need to know which data is going where. They are uncertain in determining which reports should be generated from operational systems and which one from the warehouse.
5. The users will find problems in feeding the warehouse from source systems (operational systems). In that updates have to be applied to keep data warehouse in working order.
6. Maintaining data warehouse architecture is more difficult than establishing the warehouse architecture.
7. Security policies may need to be changed depending on the user interaction with the system. Security should not be a hindrance in accessing useful information for the user of warehouse.

3.1 Data Warehouse Performance Management

The process of data warehouse performance management is similar to that of the design of a data warehouse . It is similar in that like the design and analysis phases, the procedures utilized are very different from the processes adopted in a conventional OLTP type system life cycle. In a conventional system life cycle there exists usually numerous levels of analysis and planning. In the data warehouse environment the system builders are seldom given this luxury and are required to assemble a data warehouse in a rapid manner with little time for performance analysis and capacity planning. This makes the data warehouse performance management process extremely difficult as the work loads very often cannot be predicted until finally the system is built for the first time and the data is in a production status. As a system goes into production for the first time only then may a system administrator discover there are performance problems.

3.2 Data Warehouse Maintenance

Data warehousing is becoming an increasingly important technology for information integration and data analysis . Given the dynamic nature of modern

distributed environments, both source data updates and schema changes are likely to occur autonomously and even concurrently in different data sources.
The data warehouse after its deployment needs to be treated as a production system, complete with service level agreements . Technical support for the data warehouse should constantly monitor the performance and system capacity trends and take measures to get maximum output from the system.

Six factors needed to be taken care of when dealing with ongoing data warehouse performance monitoring

1. The data warehouse grows exponentially over time in terms of size and processing requirements.
2. Capacity management estimates, even based on the most precise calculations, are most likely to be still too conservative, requiring you to consider data warehouse expansion sooner than planned.
3. Advances in technology in terms of network, hardware and software require more rapid release changes to be applied.
4. Ad hoc query access grows over time and must be carefully monitored as new and inexperienced users continue to run requests against base tables rather than summary or aggregate tables to produce totals.

An ongoing training program for business analysts, executives and decision support tool programmers keeps everyone informed as how to use the current version of the data warehouse or mart and find the information they need.

3.3 Performance Tuning Mechanisms

While the implementation of a specific phase of the data warehouse may be completed, but the data warehouse program needs to be continued . Progress monitoring needed to be continued against the agreed-on success criteria. The data warehouse team must ensure that the existing implementations remain on track and continue to address the needs of business.

Performance issues in data warehousing are centralized around access performance for running queries and incremental loading of snapshot changes from the source systems. The following six concepts can be considered for a better performance:

3.4 Network Management

If there is a heterogeneous group of platforms for the data warehouse implementation, network management is going to be one of the most demanding tasks . Not only are users coming constantly on-line, but users and equipment are invariably moving to new locations. The networking hardware is proliferating with LANs, WANs, hubs, routers, switches and multiplexers. Leaving behind all this is the next stage

3.5 Capacity Planning

Capacity planning refers to determining the required future configuration of hardware and software for a network, datacenter or web site . There are numerous capacity planning tools in the market used to monitor and analyze the performance of the current hardware and software.

3.6 Data Loading Performance

To load a data warehouse, regular loading or propagation of data from operational systems is needed . A schedule for summarizing, loading, and making the information available to the user community needs to be developed and it should be presented to the user community. For e.g. daily summary data may be available by 7 AM the next morning and weekly summary data by 8 AM Monday morning. The users should also know if and when the data was loaded.
Cleaning and transforming the data in the data staging environment prior to loading can be a great help in improved data loading performance.

1. 1 MB per second will take 41 hours to load
2. 10 MB per second will take 25 minutes to load
3. 100 MB per second will take 2.5 minutes to load

3.7 Query Management

In a data warehousing environment users queries need to be very efficiently and carefully written as some tables of the data warehouse are very huge and queries posted against these tables could days or weeks to complete.
To have an efficient query management system most of the predefined and ad hoc queries.

3.8 Software and Hardware Issues

Updates to the data warehouse are inevitable; so too will be changes to package software, hardware servers, and the supporting network infrastructure .

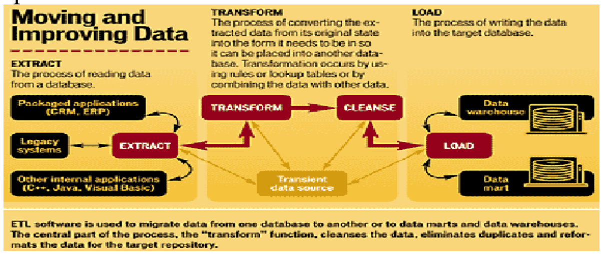
1. Installing new software releases, patches, hardware components or upgrades, and network connections (logical and physical) directly in the production environment.
2. Installing new software versions, hardware upgrades, and network improvement tasks in a

temporary test environment and migrates or reconnects to production once certification testing has concluded.

3.9 Extract, Transform and Load (ETL)

ETL is a data integration function that involves extracting data from outside sources (operational systems), transforming it to fit business needs, and ultimately loading it into a data warehouse .

Companies know they have valuable data lying around throughout their networks that needs to be moved from one place to another such as from one business application to another or to a data warehouse for analysis . The only problem is that the data lies in all sorts of heterogeneous systems, and therefore in all sorts of formats. For instance, a CRM system may define a customer in one way, while a back-end accounting system may define the same customer differently.
To solve the problem, companies use extract, transform and load (ETL) technology, which includes reading data from its source, cleaning it up and formatting it uniformly, and then writing it to the target repository to be exploited. The data used in ETL processes can come from an y source: a mainframe application, an ERP application, a CRM tool, a flat file or an Excel spreadsheet.


Moving and Improving Data
EXTRACT
TRANSFORM
LOAD
Packaged applications (CRM, ERP)
Legacy systems
Other internal applications (C++, Java, Visual Basic)
EXTRACT
TRANSFORM
CLEANSE
Transient data source
LOAD
Data warehouse
Data mart

EXTRACT
The process of reading data from a database.

TRANSFORM
The process of converting the extracted data from its original state into the form it needs to be in so it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.

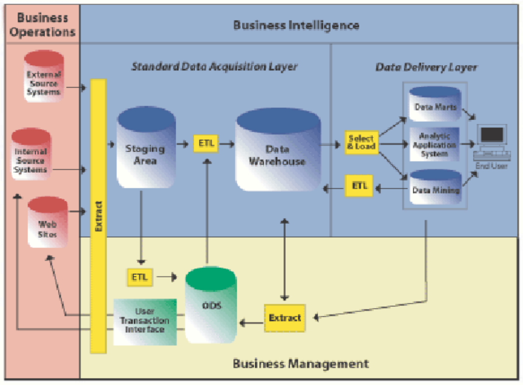
LOAD
The process of writing the data into the target database.

ETL software is used to migrate data from one database to another or to data marts and data warehouses. The central part of the process, the "transform" function, cleanses the data, eliminates duplicates and reformats the data for the target repository.

ETL function

3.9.1 A Typical ETL Process

1. The ETL architect should have a close eye on the needs and requirements of the organization. He/she must understand the overall operational environment and strategic performance requirements of the proposed system.
2. An ETL programmer should not only see his or her single-threaded set of programs.
3. The ETL process is much more than code written to move data. The ETL architect also serves as the central point for understanding the various technical standards that need to be developed if they don't already exist.
4. A key consideration for the ETL architect is to recognize the significant differences that the design and implementation methods for a business intelligence system have from an online transaction processing (OLTP) system approach. An OLTP system only changes in design when the operational process it manages changes, while BI systems must constantly adapt as business users discover new and different ways of analyzing their businesses.



ETL in Corporate Information Factory

3.9.2 Maintenance of Materialized Views

Data warehouses usually contain a very large amount of data . In this scenario it is very important to answer queries efficiently therefore we need to use highly efficient access methods and query processing techniques. It is an important physical design decision to decide which indices to build and which views to materialize. We also need to take advantage of parallel query processing to reduce query response time.

A data warehouse contains data from autonomous sources . When data in sources are updated there is a need to maintain the warehouse views in order to keep them up-to-date. This propagation of changes is commonly referred to as view maintenance and several potential policies for this have been suggested in the literature.

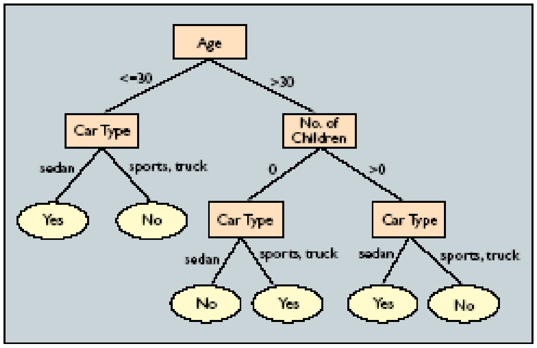
Deferred View Maintenance

As an alternative to immediate view maintenance in this technique updates to the base tables are captured in a log and applied at a later stage to the materialized views . There are further three techniques in deferred view maintenance which are:

1. Lazy: The materialized view is updated when a query accesses the view.
2. Periodic: The materialized view is updated after a certain period of time for e.g. once in a day or

during the nights etc.

3. Forced: The materialized view is refreshed after a certain number of changes have been made to the underlying tables.



6.1 Conclusion and Discussion

Data warehousing is the leading and most reliable technology used today b y companies for planning, forecasting, and management . After the evolution of the concept of data warehousing during the early 90's it was thought that this technology will grow at a very rapid pace but unfortunately it's not the reality.

A major reason for data warehouse project failures is poor maintenance. Without proper maintenance desired results are nearly impossible to attain from a data warehouse. Unlike operational systems data warehouses need a lot more maintenance and a support team of qualified professionals is needed to take care of the issues that arise after its deployment including data extraction, data loading, network management, training and communication, query management and some other related tasks.

1. Communication and Training
2. Help Desk and Problem Management
3. Network Management
4. Software and Hardware Issues
5. Extract, Transform and Load Process (ETL)

REFERENCES

WR94: Using the Data Warehouse b y W.H. Inmon and R.D. Hackathorn. 1994 John Wiley and Sons.

BS97: Data warehousing, data mining & olap authors: Alex Berson and Stephen J. Smith Publisher: Mcgraw-Hill

BSE02: A Transactional Approach to Parallel Data Warehouse Maintenance b y Bin Liu, Songting Chen, and Elke A. Rundensteiner. Worcester Polytechnic institute.2002

CM04: The computer world magazine http://www.computerworld.com/databasetopics/businessintelligence/datawarehouse/story/0,10801,89534,00.html

CT03: The ETL in a box. Claudia Imhoff and Tom Kerr. 2003 DMReview Magazine

EA99: System Analysis and Design. 2nd Edition. 1999. Elias M. Awad

EN04: Fundamentals of database systems. 4th Edition. Persons international and Addison Wesley. Ramez Elmasri and Shamkant B. Navathe

HGB01: A Benchmark Comparison of Maintenance Policies in a Data Warehouse