Research Paper
Impact Factor: 3.853
Peer Reviewed, Listed & Indexed

IJBARR
E- ISSN -2347-856X
ISSN -2348-0653

DATA WARE HOUSE MAINTENANCE DEVELOPMENT STRATEGY THROUGH EFFICIENT MAINTENANCE

Jaheeda
Associate Professor in PVKK PG College Anantapur,Andhra Pradesh .

Abstract

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations. The IT industry since mid-90.Researchers is constantly involved in finding new and improved ways for the design and development of data warehouses. But unlike traditional operational information system used for running the day to day business of an organization data warehouses require a lot more maintenance and support. 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 warehouses, 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 .As data warehousing projects are very expensive it is extremely desirable that it gives the desired results and functions smoothly. In this research study we have tried to figure out the currently available maintenance methods being used by the industry today to enhance the data warehouse performance. First we have gathered data from the books, journals, articles and the internet to see which maintenance mechanisms are available. Than we have gathered data related to data warehouse maintenance from a company using data warehouse and finally we have compared the theoretical findings with the real world findings and gave our opinion on the best possible strategies to improve data warehouse performance through efficient maintenance.

Introduction

A data warehouse is a federated repository for all the data that an enterprise's various business systems collect. The repository may be physical or logical. Data warehousing emphasizes the capture of data from diverse sources for useful analysis and access, but does not generally start from the point-of-view of the end user who may need access to specialized, sometimes local databases. The latter idea is known as the data mart... A data mart is a repository of data that is designed to serve a particular community of knowledge workers. Analysts use the data warehouse to answer unlimited variety of questions, which may be very difficult to answer in operational database. Data warehouse contains a number of databases regardless of the number. There are two approaches to data warehousing, top down and bottom up. The top down approach spins off data marts for specific groups of users after the complete data warehouse has been created. The bottom up approach builds the data marts first and then combines them into a single, all-encompassing data warehouse.

Typically, a data warehouse is housed on an enterprise mainframe. A mainframe (also known as "big iron") is a high-performance computer used for large-scale computing purposes that require greater availability and security than a smaller-scale machine can offer. Historically, mainframes have been associated with centralized rather than distributed computing, although that distinction is blurring as smaller computers become more powerful and mainframes become more multi-purpose. Server or increasingly, in the cloud. Data from various online transaction processing (OLAP) applications and other sources is selectively extracted for use by analytical applications and user queries. Of sources and volume of data. The resulted warehouse is more homogeneous compared to operational data repository. Data warehouse is often used by the large companies to analyze the data over time, and to check day to day operations. The primary goal is creating strategic planning resulting from long tern data analysis. We can create reports, projection, and business model and can forecast by these analysis. Because data stores in the data warehouse is read only and intended to provide reporting. You cannot update the data in the data warehouse by altering the records. The warehouse can be updating by adding more data from various sources and it keeping updated after a certain time period. The lifecycle of the data warehouse is continuing activity, it start form initial investigation till the requirement is met. As one phase of the data ware house is completed, other phase is started because of the new data requirements and data sources. This life cycle of the data warehouse will not end until it is valuable source of providing decision support in formation. Data warehouse is not used to keep all the data but it is used to store the necessary data for specific investigation.

Importance of the data ware house and architecture

In the current scenario of changing business conditions organizations management needs to have access to more and better information. Most organizations are now a day’s 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 their 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 1990's as large scale organizations began to need more timely data about their business, they found that traditional information systems technology was simply too slow and complex to provide relevant data efficiently and quickly. Completing reporting requests could take days or weeks using traditional reporting tools that were designed more or less to 'execute' the business rather than 'run' the business. As a cure for this problem the concept of data warehouse started as a place where relevant data could be held for completing strategic REPORTS FOR MANAGEMENT. The key here is the word 'strategic' as most executives were less concerned with the day to day operations than they were with a more overall look at the model and business functions.

In the latter half of the 20th century, there existed a large number and types of database. 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. A key idea within data warehousing is to take data from multiple platforms/technologies and place them in a common location that uses a common querying tool. In this way operational databases could be held on whatever system was most efficient for the operational business while the reporting/strategic information could be held in a common location using a common language. Data warehouses take this even a step farther by giving the data itself commonly by defining what each term means and keeping it standard. An example of this would be gender which can be referred to in many ways, but should be standardized on a data warehouse with one common way of referring to each sex. The purpose behind all these developments was to make decision sport more readily available and without affecting day to day operations. One aspect of a data warehouse that should be stressed is that it is NOT a location for ALL of a business's data but rather a location for data that is 'interesting' and 'important'. Data that is interesting will assist decision makers in making strategic decisions relative to the organization's overall mission.

Significant users of this technology include retail giants such as Wal-Mart, credit card companies such as Visa and American Express and major banks and transportation companies which include Bank of America, Royal Bank of Canada, Allied Irish Bank, United Airlines, Continental Airlines and many more. Planning the design and construction of these huge and complex information repositories have led to the development of data warehousing. Its major role remains crucial in understanding, planning scoping and delivering knowledge capital back to the enterprise in a timely and cost effective fashion.

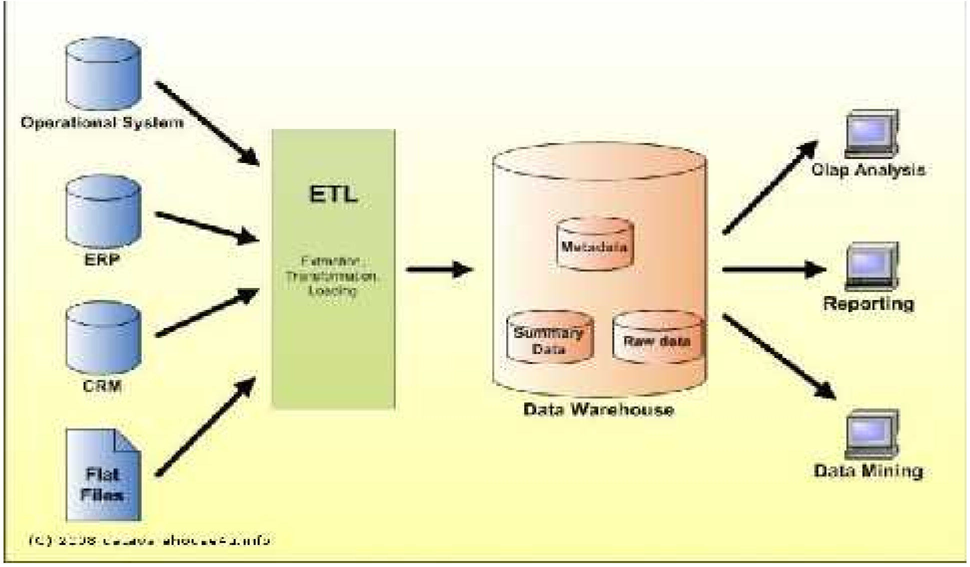


Fig**1.1 Data Warehouse in an Organization**

Another important concept that has come out of the data warehousing concept is the recognition that there are two different types of information systems in all organizations namely operational systems and information systems. Operational systems are used to perform the day to day operations of the organization. They function as a backbone to any enterprise. For e.g. order entry, inventory control, payroll, accounting etc., are all operational systems. Because of their importance the operational systems are always the first to be computerized in an enterprise. In fact most of the organizations around couldn't operate without these operations systems. On the other hand there are other functions within the organization which have to work with planning, forecasting and management. These functions are quite different from operational functions. For e.g.

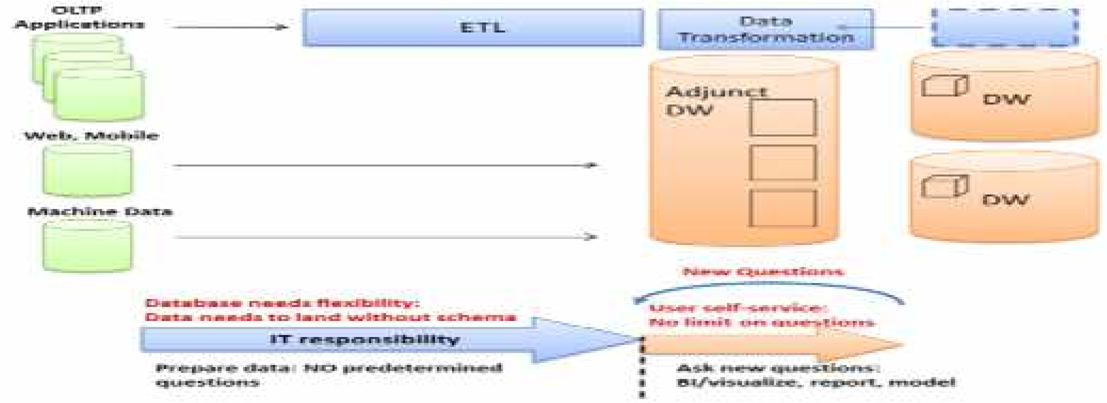
resource planning financial analysis and strategy planning etc. These functions require a lot of support from operational systems but these are actually different from operational systems. These are knowledge based systems called informational systems.

Data Warehouse Market

As a backdrop to the evaluation of various SQL-on-Hadoop products along these vectors, Gilbert identifies three key analytics usage scenarios. The first is the **core data warehouse**, a familiar concept for many tech professionals: a relatively expensive appliance-based database platform serving up highly-crated data, with the data's structure optimized for the kinds of queries the business believes it needs to run.

The second is the so-called "**data lake**" (called an "enterprise data hub" by some vendors). Here, Hadoop serves as a collecting point for disparate data sources along the full spectrum of unstructured, semi-structured and fully-structured data. Hadoop 2.0's yarn resource manager facilitates the use of a variety of analysis engines to explore the lake's data in an ad hoc fashion, and the data warehouse is relieved of this responsibility, free to serve the production queries for which it was designed and tuned.

The third scenario Gilbert identifies is one he calls the "**adjunct data warehouse**," wherein various data warehouse tasks – including ETL and reporting – are offloaded from the conventional data warehouse to Hadoop. In fact, the adjunct data warehouse can and should be used to perform these functions on data first explored in the data lake. In effect, the core data warehouse, adjunct data warehouse and Data Lake constitute a data processing hierarchy, with a corresponding hierarchy of cost. The hierarchical selection of platforms enables tasks of lower production value (though, arguably, higher business value) to be processed on cheaper platforms – yielding much higher efficiency for enterprise organizations.



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 then virtually every global 2000 company has acquired some form of data warehousing technology and is using it in some form for decision support. During the early stages of data warehouse evolution most industry professional were thinking that this technology will develop at a very rapid pace but the reality is not the same. Not very much has been accomplished market wise since its evolution. The users of data warehouse still complain about the problems of data quality, metadata management and warehouse maintenance. Users still complain that they can't get the required results from the data warehouse.

Enterprise data warehousing is a submarket within the overall data warehousing/business intelligence market. Companies considering investing in an EDW solution have matured to a point where data marts can no longer satisfy the organization's increased need for higher quality and more timely business analytics. These organizations seek a platform solution that can handle the demands of multiple subject areas and larger numbers of concurrent users, all while providing end users with the freedom to ask any question. Indeed, the warehousing market has reached a point where need, opportunity and capability have merged. Meta group believe that these forces will drive double-digit growth of the EDW market through 2008.

By the end of the 1990's most global 2000 companies had finished the major task of implementing an EFP software infrastructure. Internet has totally changed the business scenario and most of the companies have changed their existing transactional infrastructure rapidly into the web model. Missing in all this was an organization's ability to create meaning from all the transactional and sub-transactional data being captured .META Group research indicates that 77% of organizations plan to capture even more detailed business data in 2004 than was capture in 2003.

Importance of the data ware house for the it industries

A data warehouse is a database designed to enable business intelligence activities: it exists to help users understand and enhance their organization's performance. It is designed for query and analysis rather than for transaction processing, and usually contains historical data derived from transaction data, but can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources. This helps in:

- Maintaining historical records
- Analyzing the data to gain a better understanding of the business and to improve the business

In addition to a relational database, a data warehouse environment can include an extraction, transportation, transformation, and loading (ETL) solution, statistical analysis, reporting, data mining capabilities, client analysis tools, and other applications that manage the process of gathering data, transforming it into useful, actionable information, and delivering it to business users.

To achieve the goal of enhanced business intelligence, the data warehouse works with data collected from multiple sources. The source data may come from internally developed systems, purchased applications, third-party data syndicates and other sources. It may involve transactions, production, marketing, human resources and more. In today's world of big data, the data may be many billions of individual clicks on web sites or the massive data streams from sensors built into complex machinery.

Data warehouses are distinct from online transaction processing (OLTP) systems. With a data warehouse you separate analysis workload from transaction workload. Thus data warehouses are very much read-oriented systems. They have a far higher amount of data reading versus writing and updating. This enables far better analytical performance and avoids impacting your transaction systems. A data warehouse system can be optimized to consolidate data from many sources to achieve a key goal: it becomes your organization's "single source of truth". There is great value in having a consistent source of data that all users can look to; it prevents many disputes and enhances decision-making efficiency.

A data warehouse usually stores many months or years of data to support historical analysis. The data in a data warehouse is typically loaded through an extraction, transformation, and loading (ETL) process from multiple data sources. Modern data warehouses are moving toward an extract, load, and transformation (ELT) architecture in which all or most data transformation is performed on the database that hosts the data warehouse. It is important to note that defining the ETL process is a very large part of the design effort of a data warehouse. Similarly, the speed and reliability of ETL operations are the foundation of the data warehouse once it is up and running.

Users of the data warehouse perform data analyses that are often time-related. Examples include consolidation of last year's sales figures, inventory analysis, and profit by product and by customer. But time-focused or not, users want to "slice and dice" their data however they see fit and a well-designed data warehouse will be flexible enough to meet those demands. Users will sometimes need highly aggregated data, and other times they will need to drill down to details. More sophisticated analyses include trend analyses and data mining, which use existing data to forecast trends or predict futures. The data warehouse acts as the underlying engine used by middleware business intelligence environments that serve reports, dashboards and other interfaces to end users.

V **Characteristics of a Data Ware House**

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth by William Inmon:

Subject Oriented: Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a data warehouse that concentrates on sales. Using this data warehouse, you can answer questions such as "Who was our best customer for this item last year?" or "Who is likely to be our best customer next year?" This bility to define a data warehouse by subject matter, sales in this case makes the data warehouse subject oriented.

Integrated: Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile: Nonvolatile means that, once entered into the data warehouse, data should not change. This is logical because the purpose of a data warehouse is to enable you to analyze what has occurred.

Time Variant: A data warehouse's focus on change over time is what is meant by the term time variant. In order to discover trends and identify hidden patterns and relationships in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLAP) systems, where performance requirements demand that historical data be moved to an archive.

Key Characteristics of a Data Warehouse

The key characteristics of a data warehouse are as follows:

- Data is structured for simplicity of access and high-speed query performance.
- End users are time-sensitive and desire speed-of-thought response times.
- Large amounts of historical data are used.
- Queries often retrieve large amounts of data, perhaps many thousands of rows.
- Both predefined and ad hoc queries are common.
- The data load involves multiple sources and transformations.

In general, fast query performance with high data throughput is the key to a successful data warehouse.

Contrasting OLTP and Data Warehousing Environments

There are important differences between an OLTP system and a data warehouse. One major difference between the types of system is that data warehouses are not exclusively in third normal form (3NF), a type of data normalization common in OLTP environments.

Data warehouses and OLTP systems have very different requirements. Here are some examples of differences between typical data warehouses and OLTP systems:

- **Workload:** Data warehouses are designed to accommodate ad hoc queries and data analysis. You might not know the workload of your data warehouse in advance, so a data warehouse should be optimized to perform well for a wide variety of possible query and analytical operations. OLTP systems support only predefined operations. Your applications might be specifically tuned or designed to support only these operations.
- **Data modifications**: A data warehouse is updated on a regular basis by the ETL process (run nightly or weekly) using bulk data modification techniques. The end users of a data warehouse do not directly update the data warehouse except when using analytical tools, such as data mining, to make predictions with associated probabilities, assign customers to market segments, and develop customer profiles.In OLTP systems, end users routinely issue individual data modification statements to the database. The OLTP database is always up to date, and reflects the current state of each business transaction.
- **Schema design:** Data warehouses often use partially denormalized schemas to optimize query and analytical performance. OLTP systems often use fully normalized schemas to optimize update/insert/delete performance, and to guarantee data consistency.
- **Typical operation:** A typical data warehouse query scans thousands or millions of rows. For example, "Find the total sales for all customers last month."A typical OLTP operation accesses only a handful of records. For example, "Retrieve the current order for this customer."
- **Historical data**: Data warehouses usually store many months or years of data. This is to support historical analysis and reporting. OLTP systems usually store data from only a few weeks or months. The OLTP system stores only historical data as needed to successfully meet the requirements of the current transaction

Conclusion

The computational cost for running a large data ware house is very high, Accessing the information from sheer volume of data ware house is updated at regular interval. Data ware house is the leading and most reliable technology used today by companies for planning, forecasting and management like resource planning, financial forecasting and control etc., After the evaluation of the concept of data ware housing during the early 90’s it was though that this technology will grow at a very rapid pace but unfortunately it’s not the reality. A lot has been done in this field regarding design and development of data warehouses and a lot still needs to be done but one area which needs special attention from research community is data ware house maintenance. The response time significantly, Data ware housing is the leading and most reliable technology used today by companies for planning forecasting and management for resource planning. Our case study showed that services of help desk and problem management play an important role in taking valuable output from the data warehouse.

References

1. V. Markl and Bayer. processing Relational OLAP Queries with UB-Trees and Multidimensional hierarchical clustering. In proceedings of DMDW 2000,june 05-06-2000.
2. Cohen, S; Nutt, W. Serebrenik, A. Rewriting Aggregate Queries using views in 18th symposium on principles of Database Systems(PODS'99,Philadelphia,Pennsylvania,USA,may31june2)1999.
3. Albrecht, J. Hummer, W. Lehner, W. Schlesiger, Lousing semantics for Query Derivability in Data warehouse Applications, appears in proceedings of the 4th international conference of Flexible Query Answering systems.
4. Master's thesis by Bin Liu of Worcester Polytechnic Institute.
5. BS97: Data warehousing, data mining & olap authors: Alex Berson and Stephen J.Smith Publisher: Mcgraw-Hill.
6. CGO4: The evolving data warehouse market: part1. Charlie Garry 2004 Meta Delta.
7. C199: The corporate information factory. Claudia Imhoff. 1999 DMReview magazine.
8. ENO4: Fundamentals of database systems. 4th Edition. Persons international and Addison Wesley. Ramez Elmasri and Shamkant B.Navathe.
9. Shim j. Scheuermann, P. Vingalack, R. Dynamic caching of query results for decision support systems ink proceedings of the 11th international conference on scientific and statistical Database Management.
10. CTO3: The ETL in a box. Claudia Imhoff and Tom Kerr.2003 DM Review Magazine.
11. JKQ97: Algorithms for Materialized View Design in Data Warrehousing Environment, JianYang, Kmalakar Karlapalem. Qing Li. Proceedings of the 23rd VLDB conference, Athens,Greece,1997.
12. MJ96:Research Design Explained 3rd edition. Mark Mitchell and Janina Jolley(1996).
13. RK00:Data Warehouse Management Handbook by Roichard Kachur.2000 Prentice Hall.
14. RMO4: Simple strategies to improve data warehouse performance. Masters thesis by Reena Mathews of North Carolina state university,2004.
15. SU96: An Overview of Data Warehousing and OLAP Technology by Surajit Chaudhry and Umeshwar Dayal.