DATA WARE HOUSE

Bangabandhu Sheikh Mujibur Rahman Science And Technology University

Latex first project

Mamlat Biswas
A general latex hear
ID : 18ICTCSE034
23 January, 2020

Contents

Introduction to data warehouse:

A data ware house is a relation data base that is designed for query and analysis rather that for translation processing. A data ware house is a relation data base which is developed with a aim for query and analysis rather than tor transaction processing . It contains historical and commutative data derived from transaction data from single or multiple sources.Data ware house is a single verso of truth for a organization and created tor the purpose of help in decision making and forecasting .The data warehouse architecture was born in the 1980s as an architectural model designed to support the flow of data from operational systems to decision support systems. These systems require analysis of large amounts of heterogeneous data accumulated by companies over time.

In a data warehouse, data from many heterogeneous sources is extracted into a single area, transformed according to the decision support system needs and stored into the warehouse. For example, a company stores information pertaining to its employees, their salaries, developed products, customer information, sales and invoices. The CEO might want to ask a question pertaining to the latest cost-reduction measures; the answers will involve analysis of all of this data. This is a main service of the data warehouse, i.e., allowing executives to reach business decisions based on all these disparate raw data items. Thus, a data warehouse contributes to future decision making. As in the above example, a firm administrator can query warehouse data to find out the market demand of a particular product, sales data by geographical region or answers other inquiries. This provides insight about required steps to more effectively market a particular product. Unlike an operational data store, a data warehouse contains aggregate historical data, which may be analyzed to reach critical business decisions. Despite associated costs and effort, most major corporations today

use data warehouses. The data warehouse was developed in the late 1980s to meet growing demands for data analysis and information management that could not be achieved by operational systems. Because the operational systems were designed in such as way that optimizes for transactions only and number of operational or transaction systems were growing quickly across departments inside an organization that makes the data integration more difficult. This created problems of data redundancy, data integration, analysis and performance in reporting. As a result, a separate system called data warehouse is designed to solve those problems. Data warehouse system can bring data from various source systems such as relational data management systems, flat files, spreadsheets, even remote data sources outside the organization. This data then is organized in such a way that optimized for reporting purposes. User-friendly reporting tools provided by data warehouse system enable business users and decision makers to access data in the form of useful information with ease of use.

Contents

0.1 According to bill inmost

0.2 Data warehouse is

0.3 subject oriented

0.4 Integrated

0.5 Time variant

0.6 Non volatile

section Subject oriented :

It focuses on a subject rather than ongoing operations. Subject can be specific business area in an organization. For example : Sales, Marketing,Distributions . IT helps to focus on modeling and analysis of data for decision making . Image result for How data warehouse is subject oriented? A data warehouse is subject oriented as it offers information regarding a theme instead of companies' ongoing operations. These subjects can be sales, marketing, distributions, etc. A data warehouse never focuses on the ongoing operations. Instead, it put emphasis on modeling and analysis of data for decision making. Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented. You might read the Business Content chapter of Mastering the SAP Business . Information Warehouse. The book discusses these terms. However there might be a bit of overlap on what defines 'subject' and application. whether you are talking about BW or .
The book says initially BW was focused on " . . subject areas were sales . and profitability analysis." But they also go on to call these application sources.
Happy Reading.
Dave Go forth.

1 Integrated :

Integrated data from multiple data sources . Transform data from different sources into a consistent format. Must keep consistent naming conventions formats and encoding . Data integration involves combining data from several disparate sources, which are stored using various technologies and provide a unified view of the data. Data integration becomes increasingly important in cases of merging systems of two companies or consolidating applications within one company to provide a unified view of the company's data assets. The later initiative is often called a data warehouse.

Probably the most well known implementation of data integration is building an enterprise's data warehouse. The benefit of a data warehouse enables a business to perform analyses based on the data in the data warehouse. This would not be possible to do on the data available only in the source system. The reason is that the source systems may not contain corresponding data, even though the data are identically named, they may refer to different entities. At first glance, the biggest challenge is the technical implementation of integrating data from disparate often incompatible sources. However, a much bigger challenge lies in the entirety of data integration. It has to include the following phases:

Design The data integration initiative within a company must be an initiative of business, not IT. There should be a champion who understands the data assets of the enterprise and will be able to lead the discussion about the long-term data integration initiative in order to make it consistent, successful and benefitial. Analysis of the requirements (BRS), i.e. why is the data integration being done, what are the objectives and deliverables. From what systems will the data be sourced? Is all the data available to fulfill the requirements? What are the business rules? What is the support model. Analysis of the source systems, i.e. what are the options of extracting the data from the systems (update notification, incremental extracts, full extracts), what is the required/available frequency of the extracts? What is the quality of the data? Are the required data fields populated properly and consistently? Is the documentation available? What are the data volumes being processed? Who is the system owner? Any other non-functional requirements such as data processing window, system response time, estimated number of (concurrent) users, data security policy, backup policy. What is the support model for the new system? What are the SLA requirements? And last but not least, who will be the owner of the system and what is the funding of the maintenance and upgrade expenses? The results of the above steps need to be documented in form of SRS document, confirmed and signed-off by all parties which will be participating in the data integration project. Implementation Based on the BRS and SRS, a feasibility study should be performed to select the tools to implement the data integration system. Small companies and enterprises which are starting with data warehousing are faced with making a decision about the set of tools they will need to implement the solution. The larger enterprise or the enterprises which already have started other projects of

data integration are in an easier position as they already have experience and can extend the existing system and exploit the existing knowledge to implement the system more effectively. There are cases, however, when using a new, better suited platform or technology makes a system more effective compared to staying with existing company standards. For example, finding a more suitable tool which provides better scaling for future growth/expansion.

There are several organizational levels on which the integration can be performed. As we go down the level of automated integration increases.

Manual Integration or Common User Interface - users operate with all the relevant information accessing all the source systems or web page interface. No unified view of the data exists.

Application Based Integration - requires the particular applications to implement all the integration efforts. This approach is manageable only in case of very limited number of applications.

Middleware Data Integration - transfers the integration logic from particular applications to a new middleware layer. Although the integration logic is not implemented in the applications anymore, there is still a need for the applications to partially participate in the data integration.

2 Non volatile :

Data should not change once in the warehouse . Previous data is not erased when new data is added to data warehouse . Data is read only and periodically refreshed. It enables to analyses historical data and understand what and when happened . he term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization. An operational database undergoes frequent changes on a daily basis on account of the transactions that take place. Suppose a business executive wants to analyze previous feedback on any data such as a product, a supplier, or any consumer data, then the executive will have no data available to analyze becausethe previous data has been updated due to transactions. A data warehouses provides us generalized and consolidated data in multidimensional view. Along with generalized and consolidated view of data, a data warehouses also provides us Online Analytical Processing (OLAP) tools. These tools help us in interactive and effective analysis of data in a multidimensional space. This analysis results in data generalization and data mining. Data mining functions such as association, clustering, classification, prediction can be integrated with OLAP operations to enhance the interactive mining of knowledge at multiple level of abstraction. That's why data warehouse has now become an important platform for data analysis and online analytical processing. Understanding a Data Warehouse A data warehouse is a database, which is kept separate from the organization's operational database. There is no frequent updating done in a data warehouse. It possesses consolidated historical data, which helps the organization to analyze its business. A data warehouse helps executives to organize, understand, and use their data to take strategic decisions. Data warehouse systems help in the integration of diversity of application systems. A data warehouse system helps in consolidated historical data analysis. Why a Data Warehouse is Separated from Operational Databases A data warehouses is kept separate from operational databases due to the following reasons An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contract, data warehouse queries are often complex and they present a general form of data. Operational databases support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database. An operational database query allows to read and modify operations, while an OPAL query needs only read only access of stored data. An operational database maintains current data. On the other hand, a data warehouse maintains historical data. Data Warehouse Features The key features of a data warehouse are discussed below Subject Oriented A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on

the ongoing operations, rather it focuses on model ling and analysis of data for decision making. Integrated A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data. Time Variant The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.

2.1 Time

2.2 Variant

2.3 Time variant

3 Time variant :

"Time variant" means that the data warehouse is entirely contained within a time period. Another way of stating that, is that the DW is consistent within a period, meaning that the data warehouse is loaded daily, hourly, or on some other periodic basis, and does not change within that period.

Keeping in mind that these requirements were written in 1992, they are a little difficult to support some 25 years later, when real-time data warehousing is a reality, and the data warehouse may change several times per second.

Also, be careful of the definition of "non-volatile." Some people interpret it to mean that the data can never change, but this is again an outdated concept. An accumulating snapshot fact table showing a forecast shipment date for an order may well be updated many times during the life of the order. Of course, that depends on the purpose of the fact - if it was to measure volatility in forecast dates we'd keep every change, but if it is to measure the flow of an order through its life cycle, we'd just update it.

Personally, I prefer Ralph Kimball's definition, "A data warehouse is a copy of transaction data specifically structured for query and analysis." Much more clear and to the point.The time horizon for data warehouse is quite extensive compared with operational systems. The data collected in a data warehouse is recognized with a particular period and offers information from the historical point of view. It contains an element of time, explicitly or implicitly.

One such place where Datawarehouse data display time variance is in in the structure of the record key. Every primary key contained with the DW should have either implicitly or explicitly an element of time. Like the day, week month, etc.

Another aspect of time variance is that once data is inserted in the warehouse, it can't be updated or changed.

4 History

Historical data is kept in a data warehouse . Focus on change overtime is what is meant by the term variant . Helps to study trends and changes. Provides information from the historical point of view.Homepage ¿ Data Education ¿ Enterprise Information Management ¿ Information Management Articles ¿ A Brief History of the Data Warehouse A Brief History of the Data Warehouse By Keith D. Foot on April 19, 2018 Twittery be kl inked In data ware Data Warehouse (DO) stores corporate information and data from operational systems and a wide range of other data resources. Data Warehouses are designed to support the decision-making process through data collection, consolidation, analytic, and research. They can be used in analyzing a specific subject area, such as sales," and are an important part of modern Business Intelligence. The architecture for Data Warehouses was developed in the 1980 to assist in transforming data from operational systems to decision-making support systems. Normally, a Data Warehouse is part of a business mainframe server or in the Cloud. In a Data Warehouse, data from many different sources is brought to a single location and then translated into a format the Data Warehouse can process and store. For example, a business stores data about information, products, employees and their salaries, sales, and invoices. The boss may ask about the latest cost-reduction measures, and getting answers will require an analysis of all of the previously mentioned data. Unlike basic operational data storage, Data Warehouses contains aggregate historical data (highly useful data taken from a variety of sources). Punch cards were the first solution for storing computer generated data. By the 1950, punch cards were an important part of the American government and businesses. The warning "Do not fold, spindle, or mutilate originally came from punch cards. Punch cards continued to be used regularly until the mid-1980. They are still used to record the results of voting ballots and standardized tests.

magnetize storage slowly replaced punch cards starting in the 1960. Disk storage came as the next evolutionary step for data storage. Disk storage (hard drives and floppies) started becoming popular in 1964 and allowed data to be accessed directly, which was a significant improvement over the clumsier magnetic tapes.

IBM was primarily responsible for the early evolution of disk storage. They invented the floppy disk drive as well as the hard disk drive. They are also credited with several of the improvements now supporting their products. IBM began developing and manufacturing disk storage devices in 1956. In 2003, they sold their hard disks business to Hitachi.

Database Management Systems

Disk storage was quickly followed by software called a Database Management System (DBMS). In 1966, IBM came up with its own DBMS called, at the time, an Information Management System. DBMS software was designed to manage the storage on the disk" and included the following abilities: Identify the proper location of data Resolve conflicts when more than on unit of data is mapped to the same location Allow data to be deleted Find room

when stored data wont fit in a specific, limited physical location Find data quickly (which was the greatest benefit) Online Applications

In the late 1960 and early 70 commercial online applications came into play, shortly after disk storage and DBMS software became popular. Once it was realized data could be accessed directly, information began being shared between computers. As a result, there were a large number of commercial applications which could be applied to online processing. Some examples included:

Claims processing Bank teller processing Automated teller processing (Arms) Airline reservation processing Retail point of sale processing Manufacturing control processing In spite of these improvements, finding specific data could be difficult, and it was not necessarily trustworthy. The data found might be based on old information. At this time, so much data was being generated by corporations, people couldn't trust the accuracy of the data they were using.

Personal Computers and 4GL Technology

In response to this confusion and lack of trust, personal computers became viable solutions.

Personal computer technology let anyone bring their own computer to work and do processing when convenient. This led to personal computer software, and the realization that the personal computer's owner could store their "personal" data on their computer. With this change in work culture, it was thought a centralized IT department might no longer be needed.

Simultaneously, a technology called 4GL was developed and promoted. 4GL technology (developed in the 1970s through 1990) was based on the idea that programming and system development should be straightforward and anyone should be able to do it. This new technology also prompted the disintegration of centralized IT departments.

4GL technology and personal computers had the effect of freeing the end user, allowing them to take much more control of the computer system and find information quickly and efficiently. The goal of freeing end users and allowing them to access their own data was a very popular step forward. Personal computers and 4GL quickly gained popularity in the corporate environment.

But along the way, something unexpected happened. End users discovered that:

Incorrect data can be misleading. Incomplete data may not be very useful. Old data is not desirable. Multiple versions of the same data can be confusing. Data lacking documentation is questionable. Relational Databases

Relational databases became popular in the 1980s. Relational databases were significantly more user-friendly than their predecessors. Structured Query Language (SQL) is the language used by relational database management systems (RDBMS). By the late 1980s, a large number of businesses had moved from mainframe computers on to client servers. Staff members were now assigned a personal computer, and office applications (Excel, Microsoft Word, and Access) started gaining favor.

The Need for Data Warehouses

During the 1990 major cultural and technological changes were taking place.

The internet was surging in popularity. Competition had increased due to new free trade agreements, computerization, globalization, and networking. This new reality required greater business intelligence, resulting in the need for true data warehousing. During this time, the use of application systems exploded. By the year 2000, many businesses discovered that, with the expansion of databases and application systems, their systems had been badly integrated and that their data was inconsistent. They discovered they were receiving and storing lots of fragmented data. Somehow, the data needed to be integrated to provide the critical Business Information needed for decision-making in a competitive, constantly-changing global economy.

Data Warehouses were developed by businesses to consolidate the data they were taking from a variety of databases, and to help support their strategic decision-making efforts

The Use of Nosily

As Data Warehouses came into being, an accumulation of Big Data began to develop. This accumulation required the development of computers, smart phones, the Internet, and the Internet of Things to provide the data. Credit cards have also played a role, as has social media.

Facebook began using a Nosily system in 2008. NoSQL is a non relational Database Management System that uses fairly simple architecture. It is quite useful when processing Big Data. NoSQL database systems are diverse, and while SQL systems normally have more flexibility than Nosily systems, the lack (though that has changed recently) of availability in SQL gives NoSQL systems a decisive advantage. Non-relational databases (or NoSQL) use two novel concepts: horizontal scaling (the spreading of storage and work) and the elimination of the need for Structured Query Language to arrange and organize data. NoSQL databases have gradually evolved to include a wide variety of differing models. Cassandra and Hadoop are two examples of the 225+ NoSQL-style databases available.

Data Warehouse Alternatives

Data Silos can be a natural occurrence in large organizations, with each department having different goals, responsibilities, and priorities. Data silos are storage areas of fixed data which are under the control of a single department and have been separated and isolated from access by other departments for privacy and security. Data silos can also happen when departments compete instead of working together towards common goals. They are generally considered a hindrance to collaboration and efficient business practices.

A Data Mart is an area for storing data that serves a particular community or group of workers. They are storage areas with fixed data and deliberately under the control of one department within the organization.

Data Lakes use a more flexible structure for data on the way in than a Data Warehouse. Data is organized to fit the lake's database schema, and they use a more fluid approach in storing it. Data Lakes only add structure to data as it moves to the application layer. Data Lakes preserve the original structure of data and can be used as a storage and retrieval system for Big Data, which

could, theoretically, scale upward indefinitely.

Data Swamps can be the result of a poorly designed or neglected Data Lake. A Data Swamp describes the failures to document stored data correctly. This situation makes the data difficult to analyze and use efficiently. While the original data may still be there, a Data Swamp cannot recover it without the appropriate meta data for context.

A Data Cube is software that stores data in matrices of three or more dimensions. Any transformations in the data are expressed as tables and arrays of processed information. After tables have matched the rows of data strings with the columns of data types, the data cube then cross-references tables from a single data source or multiple data sources, increasing the detail of each data point. This arrangement provides researchers with the ability to find deeper insights than other techniques.

THANK YOU