A

Seminar report

On

Data Warehousing

Submitted in partial fulfillment of the requirement for the award of degree
Of Computer Science

SUBMITTED TO:

www.studymafia.org

SUBMITTED BY:

www.studymafia.org

Preface

I have made this report file on the topic **Data Warehousing;** I have tried my best to elucidate all the relevant detail to the topic to be included in the report. While in the beginning I have tried to give a general view about this topic.

My efforts and wholehearted co-corporation of each and everyone has ended on a successful note. I express my sincere gratitude to ……..who assisting me throughout the preparation of this topic. I thank him for providing me the reinforcement, confidence and most importantly the track for the topic whenever I needed it.

Introduction

Data warehouse is defined as "A subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

In this definition the data is:

• Subject-oriented as the warehouse is organized around the major subjects of the enterprise (such as customers, products, and sales) rather than major application areas (such as customer invoicing, stock control, and product sales). Date warehouse is designed to support decision making rather than application oriented data.

• Integrated because of the coming together of source data from different enterprise-wide applications systems. The source data is often inconsistent using, for example, different formats. The integrated data source must be made consistent to present a unified view of the data to the users.

• Time-variant because data in the warehouse is only accurate and valid at some point in· time or over some time interval.

• Non-volatile as the data is not updated in real time but is refreshed from on a regular basis from different data sources. New data is always added as a supplement to the database, rather than a replacement. The database continually absorbs this new data, incrementally integrating it with the previous data.

History

In the 1990's as organizations of scale began to need more timely data about their business, they found that traditional information systems technology was simply too cumbersome to provide relevant data efficiently and quickly. Completing reporting requests could take days or weeks using antiquated reporting tools that were designed more or less to 'execute' the business rather than 'run' the business.

From this idea, the data warehouse was born 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.

As with all technology, over the course of the latter half of the 20th century, we saw increased numbers 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. A key idea within data warehousing is to take data from multiple platforms/technologies (As varied as spreadsheets, DB2 databases, IDMS records, and VSAM files) 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 commonality 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).

All of this was designed to make decision support 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'. Data that is interesting will assist decision makers in making strategic decisions relative to the organization's overall mission.

Types of Data Warehouse

There are mainly three type of Data Warehouse...

1). Enterprise Data Warehouse.
2). Operational data store.
3). Data Mart.

Enterprise Data Warehouse provides a control Data Base for decision support throughout the enterprise.

Operational data store has a broad enterprise under scope but unlike a real enterprise DW. Data is refreshed in rare real time and used for routine business activity.

Data Mart is a sub part of Data Warehouse. It support a particular reason or it is design for particular lines of business such as sells, marketing or finance, or in any organization documents of a particular department will be data mart

SECURITY IN DATA WAREHOUSING

Data warehouse is an integrated repository derived from multiple source (operational and legacy) databases. The data warehouse is created by either replicating the different source data or transforming them to new representation. This process involves reading, cleaning, aggregating and storing the data in the warehouse model. The software tools are used to access the warehouse for strategic analysis, decision-making, marketing types of applications. It can be used for inventory control of shelf stock in many departmental stores.

Medical and human genome researchers can create research data that can be either marketed or used by a wide range of users. The information and access privileges in data warehouse should mimic the constraints of source data. A recent trend is to create web-based data warehouses and multiple users can create components of the warehouse and keep an environment that is open to third party access and tools. Given the opportunity, users ask for lots of data in great detail. Since source data can be expensive, its privacy and security must be assured. The idea of adaptive querying can be used to limit access after some data has been offered to the user. Based on the user profile, the access to warehouse data can be restricted or modified.

1. Replication control

Replication can be viewed in a slightly different manner than perceived in traditional literature. For example, an old copy can be considered a replica of the current copy of the data. A slightly out-of date data can be considered as a good substitute for some users. The basic idea is that either the warehouse keeps different replicas of the same items or creates them dynamically. The legitimate users get the most consistent and complete copy of data while casual users get a weak replica. Such replica may be enough to satisfy the user's need but do not provide information that can be used maliciously or breach privacy. We have formally defined the equivalence of replicas and this notion can be used to create replicas for different users. The replicas may be at one central site or can be distributed to proxies who may serve the users efficiently. In some cases the user may be given the weak replica and may be given an upgraded replica if willing to pay or deserves it.

2. Aggregation and Generalization

The concept of warehouse is based on the idea of using summaries and consolidators. This implies that source data is not available in raw form. This lends to ideas that can be used for security. Some users can get aggregates only over a large number of records where as others can be given for small data instances. The granularity of aggregation can be lowered for genuine users. The generalization idea can be used to give users high level information at first but the lower level details can be given after the security constraints are satisfied. For example, the user may be given an approximate answer initially based on some generalization over the domains of the database. Inheritance is another notion that will allow increasing capability of access for users. The users can inherit access to related data after having access to some data item.

3. Exaggeration and Misleading

These concepts can be used to mutilate the data. A view may be available to support a particular query, but the values may be overstated in the view. For security concern, quality of views may depend on the user involved and user can be given an exaggerated view of the data. For example, instead of giving any specific sales figures, views may scale up and give only exaggerated data. In certain situations warehouse data can give some misleading information; information which may be partially incorrect or difficult to verify the correctness of the information. For example, a view of a company's annual report may contain the net profit figure including the profit from sales of properties (not the actual sales of products).

4. Anonymity

Anonymity is to provide user and warehouse data privacy. A user does not know the source warehouse for his query and warehouse also does not who is the user and what particular view a user is accessing (view may be constructed from many source databases for that warehouse). Note that a user must belong to the group of registered users and similarly, a user must also get data from only legitimate warehouses. In such cases, encryption is to be used to secure the connection between the users and warehouse so that no outside user (user who has not registered with the warehouse) can access the warehouse.

The Application of Data Warehouses

The proliferation of data warehouses is highlighted by the "customer loyalty" schemes that are now run by many leading retailers and airlines. These schemes illustrate the potential of the data warehouse for "micromarketing" and profitability calculations, but there are other applications of equal value, such as:

· Stock control

· Product category management

· Basket analysis

· Fraud analysis

All of these applications offer a direct payback to the customer by facilitating the identification of areas that require attention. This payback, especially in the fields of fraud analysis and stock control, can be of high and immediate value.

Components

Operational data sources

For the DW is supplied from mainframe operational data held in first generation hierarchical and network databases, departmental data held in proprietary file systems, private data held on workstations and private serves and external systems such as the Internet, commercially available DB, or DB associated with and organization's suppliers or customers.

Operational data store

Is a repository of current and integrated operational data used for analysis. It is often structured and supplied with data in the same way as the data warehouse, but May in fact simply act as a staging area for data to be moved into the warehouse.

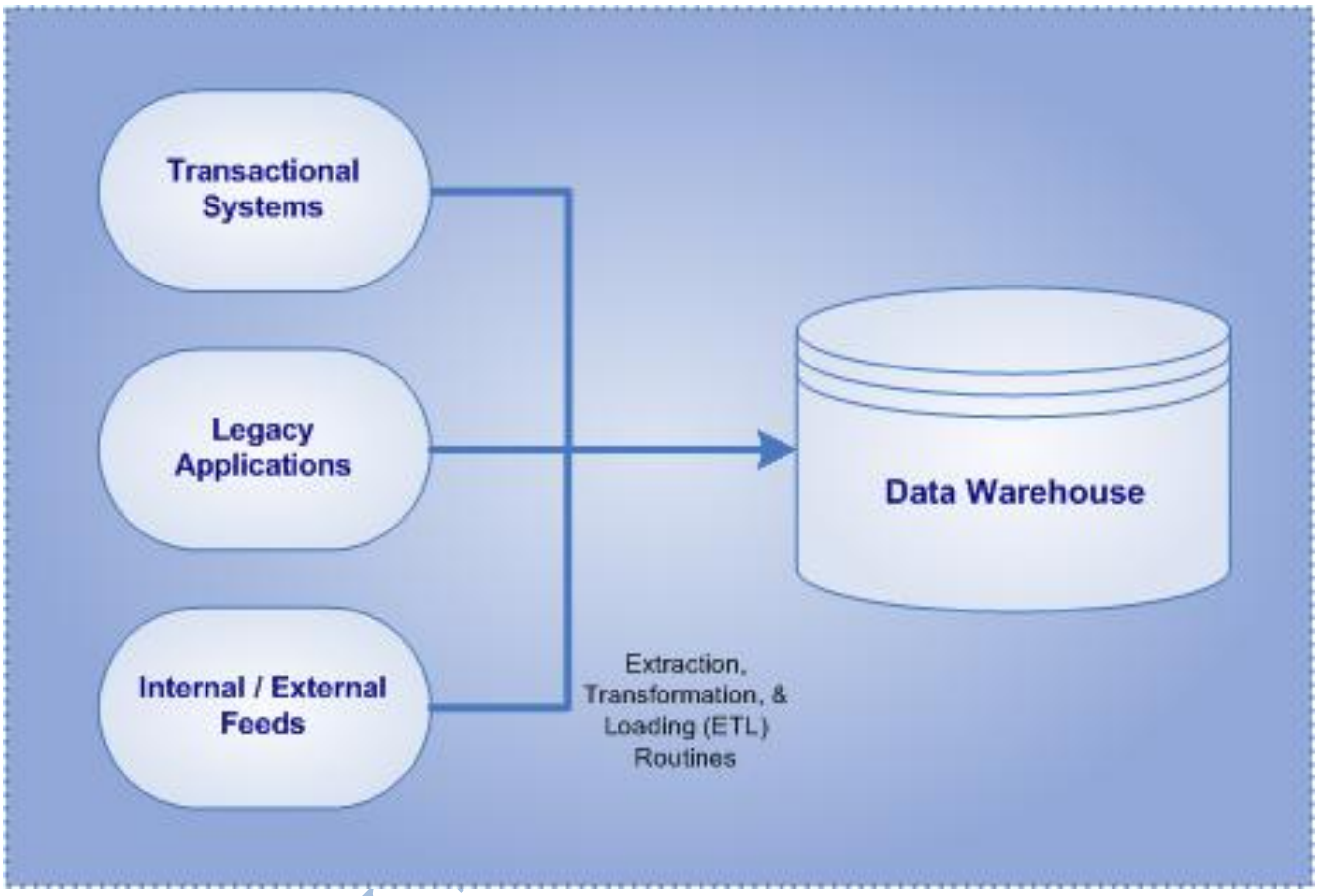
Load manager

Also called the *frontend* component, it performs all the operations associated with the extraction and loading of data into the warehouse. These operations include simple transformations of the data to prepare the data for entry into the warehouse.

Warehouse manager

Performs all the operations associated with the management of the data in the warehouse. The operations performed by this component include analysis of data to ensure consistency, transformation and merging of source data, creation of indexes and views, generation of demoralizations' and aggregations, and archiving and backing-up data.

Architecture



Benefits of Data Warehousing

The successful implementation of a data warehouse can bring major, benefits to an organization including:

• Potential high returns on investment

Implementation of data warehousing by an organization requires a huge investment typically from Rs 10 lack to 50 lacks. However, a study by the International Data Corporation (IDC) in 1996 reported that average three-year returns on investment (RO I) in data warehousing reached 401%.

• Competitive advantage

The huge returns on investment for those companies that have successfully implemented a data warehouse is evidence of the enormous competitive advantage that accompanies this technology. The competitive advantage is gained by allowing decision-makers access to data that can reveal previously unavailable, unknown, and untapped information on, for example, customers, trends, and demands.

• Increased productivity of corporate decision-makers

Data warehousing improves the productivity of corporate decision-makers by creating an integrated database of consistent, subject-oriented, historical data. It integrates data from multiple incompatible systems into a form that provides one consistent view of the organization. By transforming data into meaningful information, a data warehouse allows business managers to perform more substantive, accurate, and consistent analysis.

• More cost-effective decision-making

Data warehousing helps to reduce the overall cost of the· product· by reducing the number of channels.

Problems of Data Warehousing

The problems associated with developing and managing a data warehousing are as follows:

Underestimation of resources of data loading

Sometimes we underestimate the time required to extract, clean, and load the data into the warehouse. It may take the significant proportion of the total development time, although some tools are there which are used to reduce the time and effort spent on this process.

Hidden problems with source systems

Sometimes hidden .problems associated with the source systems feeding the data warehouse may be identified after years of being undetected. For example, when entering the details of a new property, certain fields may allow nulls which may result in staff entering incomplete property data, even when available and applicable.

Required data not captured

In some cases the required data is not captured by the source systems which may be very important for the data warehouse purpose. For example the date of registration for the property may be not used in source system but it may be very important analysis purpose.

Increased end-user demands

After satisfying some of end-users queries, requests for support from staff may increase rather than decrease. This is caused by an increasing awareness of the users on the capabilities and value of the data warehouse. Another reason for increasing demands is that once a data warehouse is online, it is often the case that the number of users and queries increase together with requests for answers to more and more complex queries.

Data homogenization

The concept of data warehouse deals with similarity of data formats between different data sources. Thus, results in to lose of some important value of the data.

High demand for resources

The data warehouse requires large amounts of data.

Data ownership

Data warehousing may change the attitude of end-users to the ownership of data. Sensitive data that owned by one department has to be loaded in data warehouse for decision making purpose.

But some time it results in to reluctance of that department because it may hesitate to share it with others.

High maintenance

Data warehouses are high maintenance systems. Any reorganization· of the business processes and the source systems may affect the data warehouse and it results high maintenance cost.

Long-duration projects

The building of a warehouse can take up to three years, which is why some organizations are reluctant in investigating in to data warehouse. Some only the historical data of a particular department is captured in the data warehouse resulting data marts. Data marts support only the requirements of a particular department and limited the functionality to that department or area only.

CONCLUSION

Since the primary task of management is effective decision making, the primary task of research, and subsequently data warehouses, is to generate accurate information for use in that decision making.

It is imperative that an organization's data warehousing strategies reflect changes in the internal and external business environment in addition to the direction in which the business is traveling.

Playing an integral role in the growth, development and success of an organization, data warehouses facilitate meaningful research which facilitates effective management.