OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of Data Mining in Genomics & Proteomics

Review Article

A Short Note on Data Warehouse

Riddhi Vora*, Sutariya Neha

Department of Computer, Shree Swami Atmanand Saraswati Institute of Technology, Surat, India

ABSTRACT

A Data Warehouse (DW) is an integrated repository of data put into a form that can be easily understood, interpreted and analyzed by the people who need to use it to make decisions. The most widely cited definition of a DW is from inmon who states that "a data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decisions."The subject-oriented property means that the data in a DW are organized around major entities of interests of an organization. Examples of subjects are customers, products, sales and vendors. This property allows users of a DW to analyze each subject in depth for tactical and strategic decision-making. The integrated property means that the data in a DW are integrated not only from all operational database systems but also some meta-data and other related external data. When data are moved from operational databases to a DW, they are extracted.

Keywords: Ethnomedicine; Ethnopharmacology; Medicinal plant; Medicinal plant park; Forest ecologic areas; Oromia; Ethiopia

INTRODUCTION

The 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 [1]. 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 because the previous data has been updated due to transactions. A data warehouses provides us generalized and consolidated data in multidimensional view [2]. 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 [3]. That's why data warehouse has now become an important platform for data analysis and online analytical processing.

LITERATURE REVIEW

Foundation of data warehousing

Data warehousing came into picture as a distinct type of computer database during the late 1980 and early 1990s. The concept of data warehousing arises to fulfil the demand of the higher management to get analytical results which normal operational database was not providing efficiently. With the improvement in technologies and higher demand from the user the concept of data warehousing has gone through several fundamental stages namely [4].

- Offline operational database
- Offline data warehouse
- Real time data warehouse
- Integrated data warehouse

Correspondence to: Riddhi Vora, Department of Computer, Shree Swami Atmanand Saraswati Institute of Technology, Surat, India; E-mail: riddhivora3839@gmail.com

Received: 06-Oct-2020, Manuscript No. JDMGP-24-6732; **Editor assigned:** 09-Oct-2020, PreQC No. JDMGP-24-6732 (PQ); **Reviewed:** 23-Oct-2020, QC No. JDMGP-24-6732; **Revised:** 01-Aug-2024, Manuscript No. JDMGP-24-6732 (R); **Published:** 29-Aug-2024, DOI: 10.4172/2153-0602.24.15.348

Citation: Vora R, Neha S (2024) Data WareHouse. J Data Mining Genomics Proteomics. 15:348.

Characteristics of data warehouse:

- Subject-oriented
- Integrated
- Time-variant
- Non-volatile

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. It also provides a simple and concise view around the specific subject by excluding data which not helpful to support the decision process [5].

Integrated

In data warehouse, integration means the establishment of a common unit of measure for all similar data from the dissimilar database. The data also needs to be stored in the datawarehouse in common and universally acceptable manner.

A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. Moreover, it must keep consistent naming conventions, format and coding [6].

This integration helps in effective analysis of data. Consistency in naming conventions, attribute measures, encoding structure etc. have to be ensured. Consider the following example (Figure 1).

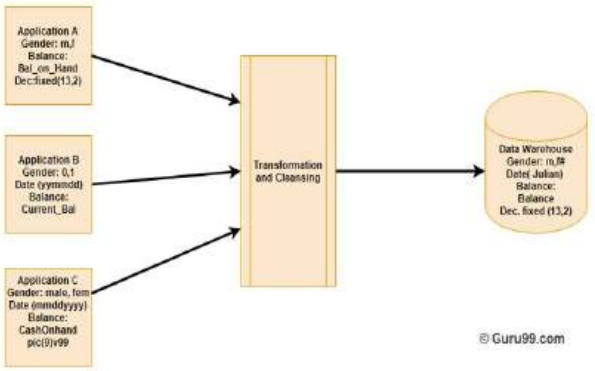

Figure 1: Data interation issues.

In the above example, there are three different application labeled A, B and C. Information stored in these applications are gender, date and balance. However, each application's data is stored different way [7].

- In application A gender field store logical values like M or F.
- In application B gender field is a numerical value.
- In application C application, gender field stored in the form of a character value.
- Same is the case with date and balance.

However, after transformation and cleaning process all this data is stored in common format in the data warehouse.

Time-variant

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 [8].

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

Non-volatile

Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it.

Data is read-only and periodically refreshed. This also helps to analyze historical data and understand what and when happened. It does not require transaction process, recovery and concurrency control mechanisms.

Activities like delete, update and insert which are performed in an operational application environment are omitted in data warehouse environment. Only two types of data operations performed in the data warehousing are (Table 1) [9].

- Data loading
- Data access

Table 1: Major differences between application and data warehouse.

Operational application	Data warehouse
Complex program must be coded to make sure that data upgrade processes maintain high integrity of the final product	This kind of issues does not happen because data update is not performed
Data is placed in a normalized form to ensure minimal redundancy	Data is not stored in normalized form
Technology needed to support issues of transactions, data recovery, rollback and resolution as its deadlock is quite complex	It offers relative simplicity in technology

DISCUSSION

Data warehouse architecture is complex as it's an information system that contains historical and commutative data from multiple sources. There are 3 approaches for constructing data-warehouse: Single tier, two tier and three tier are explained as below [10].

Single-tier architecture

The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.

Two-tier architecture

Two-layer architecture separates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.

Three-tier architecture

This is the most widely used architecture. It consists of the top, middle and bottom tier.

Bottom tier: The database of the datawarehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed and loaded into this layer using back-end tools.

Middle tier: The middle tier in data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database [11].

Top tier: The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be query tools, reporting tools, managed query tools, analysis tools and data mining tools (Figure 2).

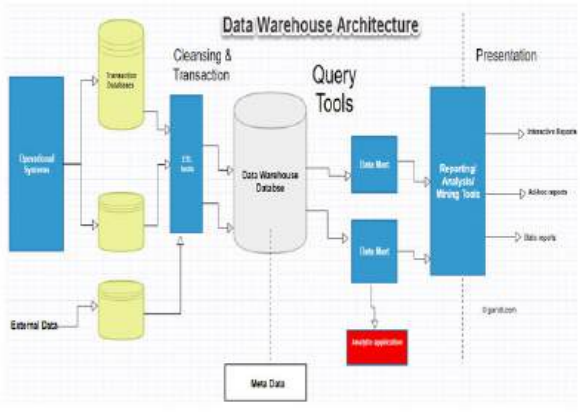

Figure 2: Data warehouse architecture.

The data warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key components to make the entire environment functional, manageable and accessible. There are mainly five components of data warehouse.

Data warehouse database

The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained by the fact that traditional RDBMS system is optimized for transactional database processing and not for data warehousing. For instance, ad-hoc query, multi-table joins, aggregates are resource intensive and slow down performance.

Hence, alternative approaches to database are used as listed below.

- In a data warehouse, relational databases are deployed in parallel to allow for scalability. Parallel relational databases also allow shared memory or shared nothing model on various multiprocessor configurations or massively parallel processors.
- New index structures are used to bypass relational table scan and improve speed.
- Use of Multidimensional Database (MDDBs) to overcome any limitations which are placed because of the relational data model. Example: Essbase from oracle.

Sourcing, acquisition, clean-up and transformation tools (ETL)

The data sourcing, transformation and migration tools are used for performing all the conversions, summarizations and all the changes needed to transform data into a unified format in the datawarehouse. They are also called Extract, Transform and Load (ETL) tools. Their functionality includes:

- Anonymize data as per regulatory stipulations
- Eliminating unwanted data in operational databases from loading into data warehouse
- Search and replace common names and definitions for data arriving from different sources
- Calculating summaries and derived data
- In case of missing data, populate them with defaults
- De-duplicated repeated data arriving from multiple datasources

These extract, transform and load tools may generate cron jobs, background jobs, cobol programs, shell scripts, etc. That regularly update data in datawarehouse. These tools are also helpful to maintain the metadata. These ETL tools have to deal with challenges of database and data heterogeneity.

Metadata

The name meta data suggests some high-level technological concept. However, it is quite simple. Metadata is data about data which defines the data warehouse. It is used for building, maintaining and managing the data warehouse. In the data warehouse architecture, meta-data plays an important role as it specifies the source, usage, values and features of data warehouse data. It also defines how data can be changed and processed. It is closely connected to the data warehouse.

For example, a line in sales database may contain: 4030 KJ732 299.90. This is a meaningless data until we consult the meta that tell us it was;

- Model number: 4030
- Sales agent ID: KJ732
- Total sales amount of $299.90

Therefore, meta data are essential ingredients in the transformation of data into knowledge. Metadata helps to answer the following questions.

- What tables, attributes and keys does the data warehouse contain?
- Where did the data come from?
- How many times do data get reloaded?
- What transformations were applied with cleansing?

Metadata can be classified into following categories.

Technical meta data: This kind of metadata contains information about warehouse which is used by data warehouse designers and administrators.

Business meta data: This kind of metadata contains detail that gives end-users a way easy to understand information stored in the data warehouse.

Query tools

One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allow users to interact with the data warehouse system. These tools fall into four different categories.

- Query and reporting tools
- Application development tools
- Data mining tools
- OLAP tools

Query and reporting tools

Query and reporting tools can be further divided into:

- Reporting tools
- Managed query tools

Reporting tools

Reporting tools can be further divided into production reporting tools and desktop report writer.

Report writers: This kind of reporting tool are tools designed for end-users for their analysis.

Production reporting: This kind of tools allows organizations to generate regular operational reports. It also supports high volume batch jobs like printing and calculating.

Some popular reporting tools are Brio, Business objects, Oracle, PowerSoft, SAS institute.

Managed query tools

This kind of access tools helps end users to resolve snags in database and SQL and database structure by inserting meta-layer between users and database.

Application development tools

Sometimes built-in graphical and analytical tools do not satisfy the analytical needs of an organization. In such cases, custom reports are developed using application development tools.

Data mining tools

Data mining is a process of discovering meaningful new correlation, pattens and trends by mining large amount data. Data mining tools are used to make this process automatic.

OLAP tools

These tools are based on concepts of a multidimensional database. It allows users to analyse the data using elaborate and complex multidimensional views.

Data warehouse models

Enterprise warehouse: It is a warehouse containing data about subject spanning the entire organization. It is usually a huge data warehouse and requires detailed business modelling. It is a data warehouse containing the data of all the subjects related to the entire organization.

Data mart: It is the subset of the enterprise data warehouse containing the data about specific subject that of value to the specific group of users. They contain information about specific subject only.

Virtual warehouse: It is built over the operational databases as a set of views. It is basically the set of views over operational database.

Problems and issues

In spite of going through huge amount research during the last decade data warehouse still have several areas to research and improve. Some of the major issues to be tackled are as follows.

Data extraction and cleaning are the first step to build a data warehouse. For any kind of database the quality of data is the most important aspect to get the desired output efficiently. Today we have number of tools available for data extraction and cleaning but they are not providing the desired efficiency. For getting the quality result it is obvious that we should have the quality data therefore extraction and cleaning of the data to get the quality data is one of keen research area for data warehouse.

Data transformation and integration is another area to be researched further as data warehouse is build up using data from heterogeneous sources therefore we should have efficient tools then available at present.

This is one of the most important tasks in data warehousing as different databases have different schemas and format and it's a prerequisite to convert them to similar format before loading into the data warehouse. The transformation of data with least error and least loss of information is still to go miles ahead.

Maintenance of a data warehouse is another aspect in which we have lot of chances to improve. We should look for some better maintenance technologies along with the software and better hardware to efficiently manage the increasing size of the data warehouse. Management of meta data should also be researched further.

Efficient retrieval of the result is the main aim of any system. In data warehouse we have several technologies available for efficient query processing but still they have to be improved a lot to achieve the required efficiency. Query processing needs to be researched further [12].

CONCLUSION

Data warehousing is the basis of automated decision support system. It has been researched a lot in the past decade but still there are many issues to be tackled in future. Performance and management are among the top research issues at present. We have identified some of the latest tools available for data warehousing and classified the tools in logical manner. The architecture of the data warehouse is also divided logically as well as a typical model of the architecture is also given. We further analysed some of the major research areas like data cleaning, data transformation, maintenance and efficient query processing. We identified major research areas in the data warehousing and the things to be done in future to achieve the best out of our data warehousing.

REFERENCES

1. Asrani D, Jain R. Designing a framework to standardize data warehouse development process for effective data warehousing practices. Int J Database Manag Syst. 2016;8(4):15-32.
2. Gupta A, Mumick IS. Maintenance of materialized views: Problems, techniques and applications. IEEE Data Eng Bull. 1995;18(2):3-18.
3. Ariyachandra T, Watson HJ. Which data warehouse architecture is most successful. Bus Intell J. 2006;11(1):4.
4. Gardner SR. Building the data warehouse. Commun ACM. 1998;41(9):52-60.
5. Golfarelli M, Rizzi S. Designing the data warehouse: Key steps and crucial issues. J Manag Inf Syst Comput Sci Appl. 1999;2(3): 88-100.
6. Winter R, Strauch B. Information requirements engineering for data warehouse systems. ACM Sympos Appl Comp. 2004;14:1359-1365.
7. Wixom BH, Watson HJ. An empirical investigation of the factors affecting data warehousing success. MIS Quart. 2001:17-41.
8. Jarke M, Jeusfeld MA, Quix C, Vassiliadis P. Architecture and quality in data warehouses: An extended repository approach. Inf Syst. 1999;24(3):229-253.
9. Sebaa A, Chikh F, Nouicer A, Tari A. Medical big data warehouse: Architecture and system design, a case study: Improving healthcare resources distribution. J Med Syst. 2018;42(4):59.
10. Ariyachandra T, Watson H. Key organizational factors in data warehouse architecture selection. Decis Support Syst. 2010;49(2): 200-212.
11. Bontempo C, Zagelow G. The IBM data warehouse architecture. Commun ACM. 1998;41(9):38-48.
12. Cabibbo L, Torlone R. An architecture for data warehousing supporting data independence and interoperability. Int J Coop Inf Syst. 2001;10(03):377-397.