BioGecko Vol 12 Issue 01 2023

ISSN NO: 2230-5807

Data Warehousing and Data Mining

Prof. Saudamini Mowade
Email: jjytode.saudamini@gmail.com

Dr. Ambedkar Institute of Management Studies & Research, Deekshabhoomi, Nagpur

Mr. Yogendra Mowade
Email: ymowade@gmail.com

Maharashtra State Power Generation Company Limited

Abstract- Data warehouses are information repositories specialized in supporting decision making. Since the de- visional process typically requires an analysis of historical trends, time and its management acquire a huge importance. In This paper shows design and implementation of data warehouse as well as its phases and its types. We recognize that, with reference to a three-tier architecture, data warehousing environment.

Keywords—data warehouse; data mart; design; MOLAP; ROLAP; data mining.

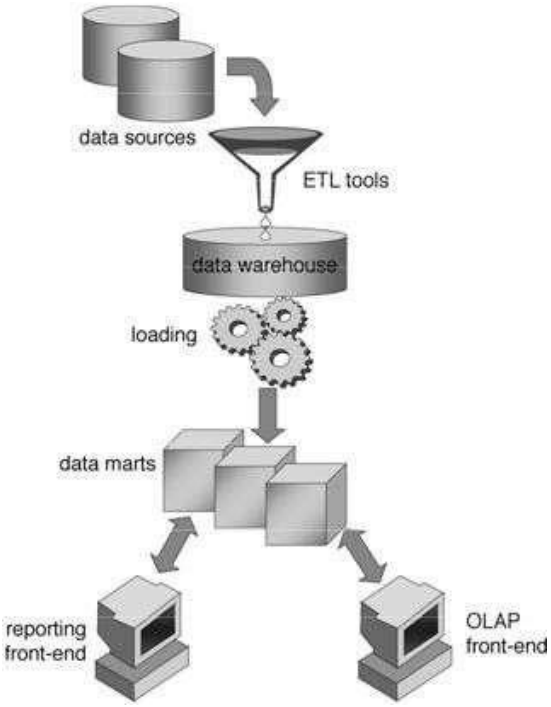
1. INTRODUCTION

Most business intelligence applications, data warehousing systems are specialized in supporting decision making. They have been rapidly spreading within the industrial world over the last decade, due to their undeniable contribution to increasing the effectiveness and efficiency of the decisional processes within business and scientific domains. This wide diffusion was supported by remarkable research results aimed at improving querying performance, at refining the quality of data, and at outlining the design process, as well as by the quick advancement of commercial tools. In the remainder of the paper, for the sake of terminological consistency, we will refer to a classic architecture for data warehousing systems, illustrated in Figure 1, that relies on three levels:

1. The data sources, that store the data used for feeding the data warehousing systems. They are mainly corporate operational databases, hosted by either relational or legacy platforms, but in some cases they may also include external web data, flat files, spreadsheet files, etc.
2. The data warehouse (also called reconciled data level, operational data store or enterprise data warehouse), a normalized operational database that stores detailed, integrated, clean and consistent data extracted from data sources and properly processed by means of ETL tools.
3. The data marts, where data taken from the data warehouse are summarized into relevant information for decision making, in the form of multidimensional cubes, to be typically queried by OLAP and reporting front-ends.

Cubes are structured according to the multidimensional model, whose key concepts are fact, measure and dimension. In the multidimensional model, events are arranged within an n-dimensional space whose axes, called dimensions of analysis, define different perspectives for their identification. Dimensions commonly are discrete, alphanumerical attributes that determine the minimum granularity for analyzing facts.

Figure 1: Three-levels architecture for a data warehousing system[6]



2. STEPS OF DATA MINING

In general, a data mining process consists of the following seven steps:

1. Identify the business problems.
2. Identify and study data sources, and select data.
3. Extract and preprocess data.
4. Mine the data, e.g., discover association rules or build predictive models.
5. Verify the mining results.
6. Deploy models in the business process.
7. Measure the return on investment (ROI).

Each data mining process is composed of a sequence of data mining operations, each implementing a data mining function or algorithm. We can categorize data mining operations into the following groups:

1. Data Understanding Operation: Access data from various sources and explore the data to become familiar with it and to \discover" the first insights.
2. Data Preprocessing Operation: Generally involves data filtering, cleaning, and transformation, to construct the final dataset for the modeling operations.
3. Data Modeling Operation: Implements the data mining algorithms, such as k-means clustering. These operations are used to build data mining models. The common modeling operations include classification, prediction, clustering, association rule, and interactive exploration such as link analysis.
4. Evaluation Operation: Used to compare and select data mining models by choosing the best one. Common operations include confusion matrix, lift chart, gain chart, cluster validation and visualization.
5. Deployment Operation: Involves deploying a datamining model to make decisions, such as using a

predictive model to predict potential customer churn or a campaign model to score customers for a target campaign.

3. DATA WAREHOUSE IMPLEMENTATION PHASES

Basic data warehouse (DW) implementation phases are :

- Current situation analysis
- Selecting data interesting for analysis, out of existing database
- Filtering and reducing data
- Extracting data into staging database
- Selecting fact table, dimensional tables and appropriate schemes
- Selecting measurements, percentages of aggregations and warehouse methods
- Creating and using the cube

The description and thorough explanation of the mentioned phases is to follow:

3.1. Current situation analysis

Computer system of FOS Student's Service Dept. was implemented at the beginning of nineties but it has been improved several times since then with the aim to adapt it to the up- to-date requests. This system fully satisfies the complex quality requests of OLTP system, but it also shows significant OLAP failures. Data are not adequately prepared for complex report forming. The system uses dBASE V database that cannot provide broad range of possibilities for creating complex reports. dBASE V does not have special tools for creating queries that are defined by the users. Design documentation is the most important in selecting of system information and data used for analysis. All vital information needed for warehouse implementation could often be found out from the design documentation of OLTP system. This phase is the most neglected one by the designers of OLTP system; therefore, their solutions do not give possibilities of good data analysis to users.

Since at this phase the possibility of realization and solution of the problem can be seen, it represents a very important phase in warehouse design. Users often know warehouse implementation.

3.2. Selecting data interesting for analysis, out of existent database

It is truly rare that the entire OLTP database is used for warehouse implementation. More frequent case is choosing the data sub-set which includes all interesting data related to the subject of the analysis. The first step in data filtering is noticing incorrect, wrongly implanted and incomplete data. After such data are located they need to be corrected if possible or eliminated from further analysis.

3.3. Filtering data interesting for analysis, out of existent database

The next step is searching for inappropriately formatted data. If such data exist, they have to be corrected and given the appropriate form. Data analysis does not need all the data but only the ones related to a certain time period, or some specific area. That is why the data reducing practice is often used.

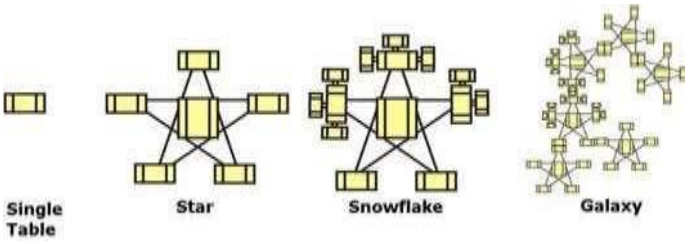
3.4. Extracting data in staging database

After the reducing and filtering of data, data are being extracted in staging database from which the data warehouse is being built. If OLAP database is designed to maintain OLAP solutions, this step can be skipped. DTS package is written in Data Transformation Services SQL Server 2000.Package writing is very important in DW implementation because packages can be arranged to function automatically so that DW system users can get fresh and prompted data.

3.5. Selecting fact table, dimensional tables and appropriate schemas

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for on-line transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates online data analysis. Figure 2 shows the schemas that are used in implementation of Data warehouse system.

Figure 2: Data warehouse schema based on [7].



The simplest scheme is a single table scheme, which consists of redundant fact table. The most common modeling paradigm according to [9] is star schema, in which the data warehouse contains a large central fact table containing the bulk of data, with no redundancy, and a set of smaller attendant tables (dimension tables), one for each dimension. Snowflake schema is a variant of star schema model, where some dimension tables are normalized, causing thereby further splitting the data into additional tables. Galaxy schema is the most sophisticated one, which contains star and snowflake schemas.

3.6. Selecting measurements, percent of aggregations and warehouse modes

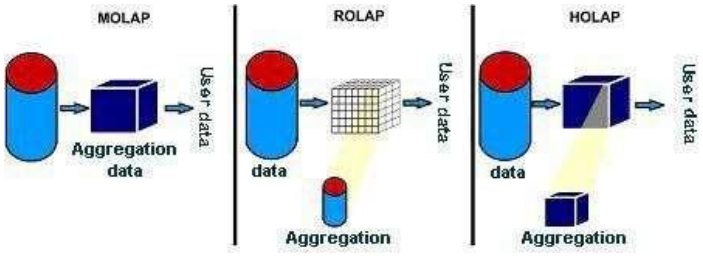
The next step in designing data warehouse is selecting measurements. In this case, two measurements can be seen: total number of passed exams and average mark achieved in passed exams. In the data warehouse implementation very often appears the need for calculated measurements that are attained from various arithmetic operations with other measurements. Furthermore, this system uses the average that has been calculated as the ratio of the total mark achieved on passed exams and the number of passed exams.

Data warehouse solutions use aggregations as already prepared results in user queries and through them they solve the queries very fast. The selection of an optimal percentage of aggregation is not simple for the designer of the OLAP system. The increasing of the percentage of aggregated data speeds up the user-defined queries, but it also increases also the memory space used.

The most important factors that can have an impact on the storing mode are:

The size of the OLAP base, the capacity of the storage facilities and the frequency of data accessing. Manners of storing are: ROLAP (RELATIONAL OLAP),HOLAP (HYBRID OLAP) and MOLAP (MULTIDIMENSIONAL OLAP). which is shown on fig3.ROLAP stores data and aggregation into a relational system and takes at least disc space, but has the worst performances. HOLAP stores the data into a relational system and the aggregations in a multidimensional cube. It takes a little more space then ROLAP does, but it has better performances. MOLAP stores data and aggregations in a multidimensional cube, takes a lot of space, but has the best performances since very complex queries will be used in analysis it is rational to use MOLAP.

Figure 3: Modes of data storing based on [8].



3.7 Creating and using the cube

The cube is being created on either client or server computer. Fundamental factors that influence the choice of

the place for cube's storehouse are: size of the cube, number of the cube users, performances of the clients and server's computers and throughput of the system. The created cube can be used by the support of various clients tools.

Figure 4: Multidimensional cube

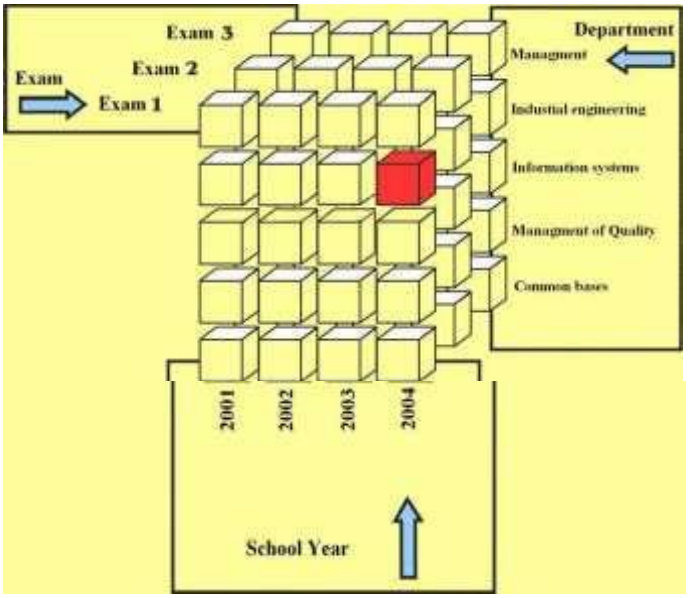


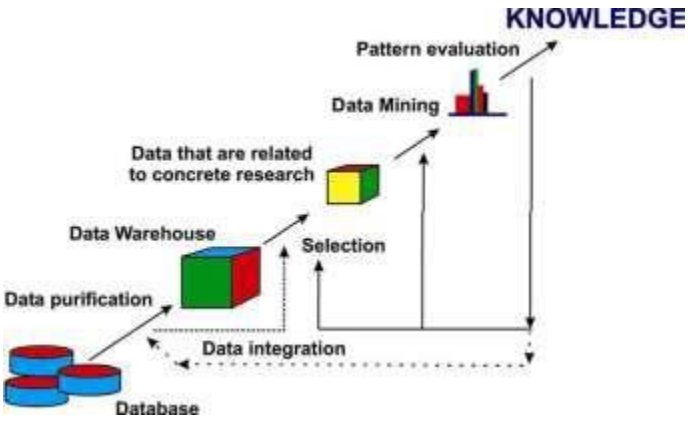
Figure 4 shows a three-dimensional cube that represents average grades by department, Exam and school year. In comparison with OLTP systems that have to calculate the average grade on each request, data warehouse systems have results prepared in advance and stored in multidimensional format.

4. FROM DATA WAREHOUSE TO DATA MINING

The previous part of the paper elaborates the designing methodology and development of data warehouse on a certain business system. In order to make data warehouse more useful it is necessary to choose adequate data mining algorithms. Those algorithms are described further in the paper for the purpose of describing the procedure of transforming the data into business information i.e. into discovered patterns that improve decision making process. DM is a set of methods for data analysis, created with the aim to find out specific dependence, relations and rules related to data and making them out in the new, higher-level quality information [2]. As distinguished from the data warehouse, which has unique data approach, DM gives results that show relations and interdependence of data. Mentioned dependences are mostly based on various mathematical and statistic relations [3]. Figure 9 represents the process of knowledge data discovery.

Data for concrete research are collected from internal database system of Student's Service Dept., and external bases in the form of various technological documents, decisions, reports, lists, etc. After performed selection of various data for analysis a DM method is applied, leading to the appropriate rules of behavior and appropriate patterns. Knowledge of observed features is presented at the discovered pattern. DM is known in literature as the "extraction of knowledge", "pattern analysis", "data archaeology" [3].

Figure 5: Process of knowledge data discovery based on[9]



5. OPEN SOURCE SYSTEMS

Open source software provides users with the freedom to run, copy, distribute, study, change and improve the software. To adopt open source software (or in fact any software) we must understand its license and the limitations that it places on us. Unlike closed source licenses which aim to limit your rights, open source software aims to give you the right to do whatever you please with the software. The common open sources licenses include the GPL, LGPL, BSD, NPL, and MPL. Bruce has discussed each license with suggestions for choosing a proper license [12].

In the past decades, open source products such as GNU/Linux, Apache, BSD, MySQL have achieved great success, clearly demonstrating that open source software can be as robust, or even more robust, than commercial and closed source software.

In the next subsection we discuss some important features to categorize open source data mining systems.

Important Features of Open Source Data Mining Systems

Open source systems are diverse in design and implementation. Although developed for data mining, they are very different in many aspects. To understand the characteristics of these diverse open source data mining systems and to evaluate them, we look into the following important features:

- Ability to access various data sources.

Data comes from databases, data warehouses, and flat files in different formats. A good system will easily access different data sources.

- Data preprocessing capability.

Preprocessing occupies a large proportion of the time in a data mining process [13]. Data preparation is often the key to solving the problem. A good system should provide various data preprocessing functions tasks easily and efficiently.[14].

- Integration of different techniques.

There is no single best technique suitable for all data mining problems. A good data mining system will integrate different techniques (preprocessing functions and modeling algorithms), providing easy access to a wide range of different techniques for different problems.

- Ability to operate on large datasets.

Commercial data mining system, such as SAS Enterprise, can operate on very large datasets. This is also very important for open source data mining systems, so scalability is a key characteristic.

- Good data and model visualization.

Experts and novices alike need to investigate the data and understand the models created.

- Extensibility.

With new techniques and algorithms it is very important for open source data mining systems to provide architecture that allows incorporation of new methods with little effort. Good extensibility means easy integration of new methods.

- Interoperability with other systems

Open standards means that systems (whether open or closed source) can interoperate. Interoperability includes data and model exchange. A good system will provide support of canonical standards, such as CWM [2] and PMML [3].

- Active development community.

An active development community will make sure that the system is maintained and updated regularly. These features categorize open source data mining systems. In the following, we investigate commonly used open source data mining systems with respect to these features.

6. CONCLUSION

In this paper we have presented important features for open source data mining systems. We present the four following important points for open source data mining systems to gain greater success in deployment:

1. Supporting various data sources

A good open source data mining system should support most commonly used data sources, such as the open source and commercial databases, csv files, and user defined formats.

2. Providing high performance data mining

Most open source data mining system can't operate on large volumes of data. To offer high performance data mining we need to either rewrites the algorithms (e.g. parallel and distributed algorithms) or more simply to improve the hardware on which the software is running.

3. Proving more domain-specific techniques

Most data mining system integrate many algorithms at the whim of the researchers, rather than for the benefit of business. We need to better identify algorithms that match the data to be processed. One approach will be to provide domain-specific techniques based on a generic platform.

4. Better support for business application

Real business application are complex, placing many demands on the data mining system. Open source data mining systems need to in improve scalability, reliability, recoverability, and security [1].

7. REFERENCES

1. Kleissner, C.: Data mining for the enterprise. In: In Proceeding of the 31st Annual Hawaii International Conference on System Science. (1998) 295{304}
2. Object Management Group: Common warehouse met model (cwm) (2007) Web-site http://www.omg.org/cwm/
3. Data Mining Group: Predictive model markup language (pmml) (2005)
4. Information Technology and Systems Center (ITSC) at the University of Alabama in Huntsville: Algorithm development and mining system (2005) Website: http: //datamining.itsc.uah.edu/adam/.
5. HIT-HKU BI Lab: Alphaminer 2.0 (2006) Website:http://bi.hitsz.edu.cn/

6. International Journal of Data Warehousing & Mining,5(1), 1-17, January-March 2009
7. Vidette, P., Building a Data Warehouse for Decision Support, Prentice Hall, 1996.
8. Lory, O., and Crandall, M., Programmers Guide for Microsoft SQL Server 2000, Microsoft Press, 2001
9. Jiwei, H., and Micheline, K., Data Mining: Concepts and Techniques, Simon Fraser University, 2001.
10. Krulj, D., "Design and implementation of datawarehouse systems", M Sc. Thesis, Faculty of Organizational Sciences, Belgrade, 2003.
11. Krulj, D., Suknović, M., Čupić, M., Martić, M., and Vujnović, T., "Design and development of OLAP system FOS student service", INFOFEST, Budva, 2002.
12. Seidman, C., Data Mining with Microsoft SQL Server 2000,Microsoft Press, 2001.
13. Grossman, R., Kamath, C., Kegelmeyer, P., Kumar, V.,Namburu, R.: Data Mining for Scientific and Engineering Applications. Kluwer Academic Publishers (2001).
14. National Natural Science Foundation of China(NSFC) under grants No.6060306 International Journal of Data Warehousing & Mining,5(1), 1-17,January-March 2000.