DESIDOC Journal of Library & Information Technology, Vol. 29, No. 2, March 2009, pp. 61-66

Advanced Applications of Data Warehousing Using 3-tier Architecture

Praveen Sharma

Deptt of M.B.A (Systems) and Engg. Chemistry
Venkteshwar Institute of Technology, Indore-453 331
E-mail: praveen.s.sharma@email.com

ABSTRACT

Organisations be it industry or business or even educational institutes, need to improve their information inventory system so as to survive in the competitive environment. The organisations have to increase their efficiency and effectiveness in maintaining the cycle of activities, in their planning, decision-making processes, and analytical needs. There are several ways to acquire this goal; one of it is with data mining which is able to make a prediction using existing data in their database in order to forecast future demand. In addition, with data mining they would be able to determine which activity is more important and what trend is prevailing. An information system, which is based on both World Wide Web technology and a 3-tiered architecture, is proposed herein to meet the above requirements. This paper is an attempt to provide the initial concept about data mining model that most likely will be used in various department including libraries of the teaching institutes. The initial concepts covered by the paper are the appropriate data warehouse schema; data mining tasks and techniques that are best suited, and applications.

Keywords: OLAP, OLTP, data warehouse, dashboard, data mart

1. INTRODUCTION

Data warehousing is an algorithm and a tool to collect the data from different sources and to store it in a single repository to facilitate the decision-making process. According to W. H. Inmon[1], a leading architect in the construction of data warehouse systems, a data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.

This short, but comprehensive definition presents the major features of a data warehouse. The four keywords—subject-oriented, integrated, time-variant, and non-volatile—distinguish data warehouses from other data repository systems such as relational database systems, transaction processing systems, and file systems.

Subject-oriented: A data warehouse is organised around major subjects, such as customer, vendor, product, and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organisation, a data warehouse focuses on the modelling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision-support process.

Integrated: A data warehouse is usually constructed by integrating multiple heterogeneous sources such as relational databases and online transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

Time variant: Data are stored to provide information from a historical perspective (e.g., the past 5-10 years).Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

Non-volatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

Received on 1 September 2008

Data warehousing provides an interesting alternative to the traditional approach of heterogeneous database integration described above. Rather than using a query-driven approach, data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis. Major libraries have large collections and circulation. Managing libraries electronically has resulted in the creation and management of large library databases. The interconnection of libraries and sharing resources across libraries has resulted in the management of very large databases. Most large and/or multinational industries worldwide have exploited such opportunities by applying data warehouse technology to their data repositories to discover knowledge that had helped them to gain competitive advantage through decision making. The same can be done for libraries using the available large databases. This paper identifies how the data warehouse technology could assist libraries to discover knowledge and improve services.

2. OPERATIONAL DATABASE SYSTEM VERSUS DATA WAREHOUSES

Since most people are familiar with commercial relational database systems, it is easy to understand what a data warehouse is by comparing these two kinds of systems. The major task of online operational database systems is to perform online transaction and query processing. These systems are called online transaction processing (OLTP) systems, and cover most of the day-to-day operations of an organisation such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

Data warehouse systems, on the other hand, serve users or\knowledge workers in the role of data analysis and decision making. Such systems can organise and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as online analytical processing (OLAP) systems. The major distinguishing features between OLTP and OLAP are:

(i) Users and system orientation: An OLTP system is customer-oriented and used for normal transaction like issue of book, and query processing by librarians, teachers, and information technology professionals. An OLAP system is more of planning and decision-making oriented and is primarily used by knowledge workers including managers, executives, and analysts.

(ii) Data contents: An OLTP system manages current data that, typically, are too detailed and used to run the day-to-day business of the institution library. An OLAP system manages large amounts of historical data, provides facilities for summarisation and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for use in informed decision making.

(iii) Database design: An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design.

(iv) View: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organisations. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organisation. OLAP systems also deal with information that originates from different organisations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

(v) Access patterns: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations (since most data warehouses store historical rather than up-to-date information), although many could be complex queries. Other features which distinguish between OLTP and OLAP systems include database size, frequency of operations, and performance metrics.

3. THE 3-TIER ARCHITECTURE

The data warehousing has three-tier architecture (Fig. 1). The first-tier is known as the extraction and transformation tier. The second-tier is known as middle or connective tier, and the third-tier is known as data access and retrieval tier.

3.1 Extraction and Transformation Tier

The extraction is the process of refining the data that is collected from the different sources like internal database of the organisation, external databases from various departments of the institute, other leading educational libraries in the city, etc. Two methods can be used for the extraction of the data from sources, viz., bulk extraction and change-based extraction.

The entire process of extracting data from multiple sources, transforming it into a unique standard format and finally the loading into the warehouse is referred as

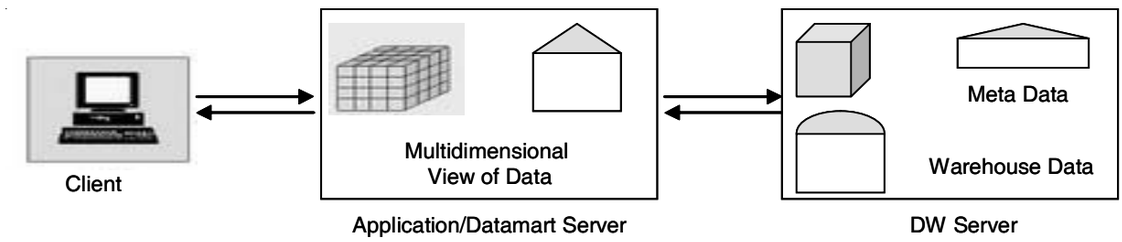

Figure 1. Three-tier architecture of data warehouse.

extraction, transformation and loading (ETL) process. The transformation tolls helps in transforming the data in to a particular format after extraction and before loading it in to the warehouse.

Client

- GUI/presentation logic
- Query specification
- Data analysis
- Report formatting
- Data access

Application/Data Mart Server

- Summarising
- Filtering
- Meta data
- Multidimensional view
- Data access

DW Server

- Data logic
- Data services
- Metadata
- File services

3.2 Connective Tier

Data source rules and the programs applied for the transformation, the data structure and contents are all members of the meta data, which is known as data about data. In the middle tier within the meta data, data warehouse is a single and central repository of the enterprise wide data of all possible subject areas.

3.3 Data Access and Retrieval Tier

OLAP tool allows the user to generate queries against the warehouse database while reporting tool helps to produce canned, graphic intensive, sophisticated reports. The non-trivial extraction of actionable knowledge from datasets is known as data mining.

4. APPLICATIONS

The implementation of projects using inappropriate technology has become a common practice. More often than not, these projects end up with for lack of a better term, confused systems as they meet some needs but not others. The thing that keeps these systems alive is the substantial investments of time and money. Even than, after a while the money stops playing a role and these systems end up ignored or replaced. To ensure that data warehousing endeavours to not suffer the same fate, it's important to recognise where it makes sense to implement them. The type of database architecture that has recently emerged is data warehouse, which is a repository of multiple heterogeneous data sources, organised under a unified schema at a single site in order to facilitate management decision-making[2,3]. The data warehousing makes sense in two primary areas: operational, and business intelligence applications.

4.1 Operational Applications

Contrary to popular belief, successful data warehouses are utilised for operational applications. Library cards and usage reporting applications seem to be the first operational applications to take advantage of real-time data warehousing at places such as digital libraries. To keep up with the library cards and usage volumes, the libraries should focus from night batching jobs to real-time or near real-time data processing. Data warehouses are the natural choice for this type of processing as aggregation and fast retrieval of large amounts of data is what they do best.

Can a library grow up to a terabyte of data? In most cases it does not as even the storage space of computers used by most libraries are not in this range. Let us perform a calculation to determine the storage requirements for a library. If a catalogue consists of 200 characters it would use 200 kilobytes of data. If there are 1 million items then the storage requirement would be 200 million kB or 200 GB. If a library has 50,000 members with each taking 200 kilobytes then the storage requirement would be 10 GB. If a circulation record is using 40 bytes per item and if 1000 item borrowing happens per day then the storage requirement would be 40 kB. Over a year this would accumulate to about 10 GB. Thus the total storage

requirement to manage electronic transactions will not fall under the very large database category. Digital libraries, also called electronic libraries, are being widely adopted across many libraries and thus have moved from relatively few people's research interests to a wider application and use. Digital libraries have integrated different information sources and increased the use of information. Functions of a library have grown beyond maintaining books, magazines, and newspapers. Many libraries also provide CD/video lending, and online searching, reservations and browsing e-journals. Certain universities and libraries have even moved beyond this level and provide full-text books, multimedia manuscripts, and periodicals[4]. Newspapers and technical documents are also available on the web along with their print edition. Education is now moving towards electronic learning environment[5] and digital libraries will play a major role in achieving it. A recent development in the field is the step by IIM, Indore, to interconnect all the libraries in Indore. Following are among the benefits digital libraries would bring to the libraries in particular and society in general[6]:

- There will not be any boundaries in the distribution and dissemination of information.
- The performance of a library would increase immensely while accessibility to all kinds of items being provided through a single workstation located anywhere.
- Although dealing with more people and information, the administrative overheads will be very minimal as computers do most of the tasks without any human intervention.

Data of a digital library should be organised in a manner in which it could be analysed later. Classification of data according to identified characteristics is one way of achieving it. Classification will allow grouping of discrete values such as by subject of a book. However, estimation would be required to deal with continuous values such as age groups of borrowers. Classified and estimated data could be used to predict future behaviours and decision making could be done accordingly. Membership characteristics such as disability, ethnic group income group, and social characteristics could help to determine affinity groups. Services offered could target such groups. For example, reduced membership fees or special facilities may be offered to different affinity groups to ensure that they are not neglected. Clustering is another technique used to effectively organise information by segmenting a heterogeneous population into a number of more homogeneous sub-groups or clusters. Clustering differs from classification as it is based on self-similarity than predefined classes. For example, one may group books based on author or publisher due to the high probability of such books falling into the same group. Referencing of the required information form the large amount of redundant information available in a library can be reduced using mining techniques. Searching through classification of content of the library and acquisition of books through data mining knowledge are among the main benefits that a library would gain through the management of large databases and data warehousing.

A library application would manage catalogues using the database technology. Such applications are referred to as e-catalogue. The entries in an e-catalogue are reusable since a user could search for a catalogue entry not only by author name but also by any other data field such as title and publisher. Library application packages such as LibSys and CDS/ISIS have enabled effective management of e-catalogues. Other operational applications include inventory management and service provisioning, to mention a few. Many businesses have moved to just-in-time inventory. Once provisioned for some service, the billing clock starts ticking, so tracking when the service was provisioned is significant. The temporal element is the common denominator for operational applications that can benefit from real-time data warehousing. The complex analysis in the library system can be easily done by using the multi-dimensional data. To facilitate complex analyses and visualisation, the data in a warehouse is typically modelled multi-dimensionally. For example, in a library data warehouse, time of borrowing, borrower's district, age group and book category might be some of the dimensions of interest (Fig. 2). Often, these dimensions are hierarchical, e.g., time of borrowing may be organised as a day-month-quarter-year hierarchy.

4.2 Business Intelligence Applications

A decision support system (DSS) is a tool used to improve the process of decision-making in complex

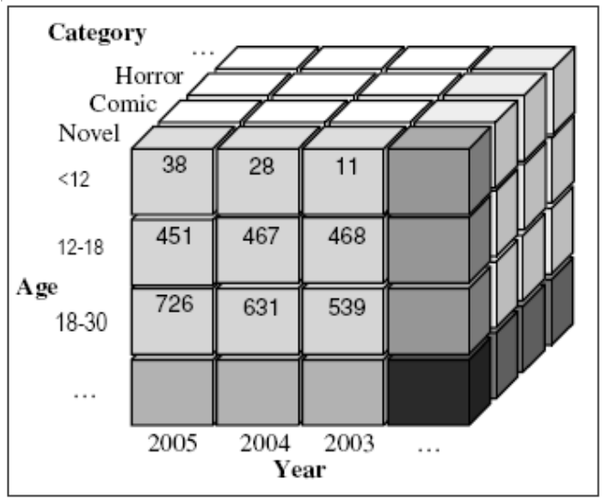

Figure 2. A Multi-dimensional data cube.

systems. A DSS can range from a system that answer simple queries and allows a subsequent decision to be made, to a system that employ artificial intelligence and provides detailed querying across a spectrum of related datasets. Amongst the most important application areas of DSS are those complicated systems that directly "answer" questions, in particular high level "what-if" scenario modeling. Over the last decade there was a transition to decision support using data warehouses[7]. The data warehouse environment is more controlled and therefore more reliable for decision support than the previous methods. The data warehouse environment supports the entire decision support requirements by providing high-quality information, made available by accurate and effective cleaning routines and using consistent and valid data transformation rules and documented pre-summarisation of data values. It contains one single source of accurate, reliable information that can be used for analysis. Most large and/or multinational industries worldwide have already applied data warehouse technology to their data repositories and have discovered knowledge to gain competitive advantage through decision making. Thus researchers in libraries are trying to do the same to discover similar opportunities[8-11]. For instance, a decision like "on which books should a library invest to serve their members better?" could be made after such discovery.

One of the current trends in business intelligence is the implementation of corporate dashboards. The sophistication of corporate dashboards is quite broad and ranges from a report that lists several key performance indicators (KPIs) to dashboards that mimic the look and feel of automobile dashboards right down to the round gauges with needles that fluctuate in real time. Timeliness of the dashboards is driven by the data freshness in the data warehouse. Dashboards that display real-time data activity typically have a real-time data warehouse behind them. The kind of information reported by these dashboards is as varied as the dashboards themselves. Financial information typically takes first dibs followed by trend and operational information. Financial information includes up-to-date revenue, current exchange and interest rates, and stock prices to mention a few KPIs. Trend reports typically display KPIs across some pre-defined timeframe such as rolling 24 hr view, while operational KPIs can include equipment utilisation rates.

Customer relationship management (CRM) business intelligence is an area that has been getting quite a bit of attention and where real-time capability is a value- add, especially in a web environment. The Internet is definitely an area that can and should reap the benefits of real-time data warehousing. Clickstream data generates substantial transaction volumes that, in popular sites, can easily eclipse credit card transaction volumes. These data volumes alone justify having real-time data warehousing. From a business intelligence perspective, for clickstream data to make sense, it has to be aggregated into predefined buckets to allow for click activity analysis. Once again, data warehousing is the natural choice for this type of processing.

5. CONCLUSION

Digital libraries have emerged over the last decade and are being used beyond the research community and selected institutions. Libraries should prepare to exploit these digital collections for decision-making and provide services to suit the digital society. This can be facilitated by using the three tier architecture of data warehousing. This three-tier approach has several advantages as the data mining services can control the number of connections to the warehouse as well as the number of statements currently executed by the database system. The middle tier can control the number and kind of data mining tasks that are processed in parallel. This enables the system to incense the usage of system resources for data mining purposes, especially bandwidth and CPU cycles. Second, the system can service users according to their priority and membership in user groups. This includes restricted access to data mining tables as well as user specific response behavior. Third, a wide range of optimisation strategies can be realised. The tasks of the data mining services can be distributed over the client and the middle tier. The middle tier can exploit parallelism by parallel processing on the middle tier hardware and parallel connections to the database layer. Additionally, the data mining services can re-use the outcome of data mining sessions and pre-compute common intermediate results. In summary, the main advantage of three-tier architecture is that mining can be done in a controlled and manageable way for multiple users. Data warehousing systems have become a key component of information technology architecture. A flexible enterprise data warehouse strategy can yield significant benefits for a long period.

REFERENCES

1. Inmon, W.H. Buliding the dataware house. QED Technical Publishing Group, Wellesley, Massachutusetts.

2. Chaudhuri, S. & Dayal U. An overview of data warehousing and OLAP technology. SIGMOD Record 26:1, 1997, March 65-74.

3. Han, J. & Kamber, M. Data mining concepts and techniques, edited by Morgan Kaufmann, V. Harinarayan, A. Rajaraman, J.D.Ulman. *In* Implementing Data Cubes Efficiently. Proceedings of SIGMOD Record, **25(**2), 205-16.

4. Chen, H. The Illinois digital library initiative project: Federating repositories and semantic research. Available at http://ai.bpa.arizona.edu/hchen/docs/DLI/

5. Wikramanayake, G.N. Impact of digital technology on education, *In* Proceedings of 24th National Information Technology Conference, Computer Society of Sri Lanka, Colombo, Sri Lanka, 15-16 August 2005, 82-91.

6. Source Book on Digital Libraries, TR 93-35, edited by E. A. Fox. Dept. of Computer Science, Virginia Tech University, 1993. Available at http://fox.cs.vt. edu/DigitalLibrary/DLSB.pdf.

7. Inmon, W.H. Building the data warehouse, 3rd Edition, Wiley. Needamangala, A. A Library Decision Support System Built on Data Warehousing and Data Mining Concepts and Techniques, Master of Science Thesis, University of Florida, 2002.

8. Guenther, K. Applying data mining principles to library data collection. *Computers in Libraries*, **20**(4), 60-63.

9. Baruque, C.B. & Melo, R.N. Developing digital libraries using data warehousing and data mining techniques, Available at http://www.pgl.ufl.edu/events/pgl2/CBaruque/digital_libraries.pdf

10. Dwivedi, R.K. & Bajpai, R.P. Use of data mining in the field of library and information science: An overview. 2004.

11. Prakash, K.; Chand, P. & Gohel, U. Application of data mining in library and information services. *In* 2nd Convention PLANNER-2004, Manipur Uni., Imphal, 4-5 November 2004.