Data Mining and Warehousing Approaches on School Smart System: A Conceptual Framework

Janson Luke Ong Wai Kit[1] Subarmaniam Kannan[2] Vijanth .S Asirvadam[3]

[1]Faculty of Information Science and Technology
Multimedia University Melaka Campus Jalan Ayer Keroh Lama, 75450, Melaka , Malaysia
E-mail: luke.wai.kit04@mmu.edu.my

[2] Faculty of Information Science and Technology
Multimedia University Melaka Campus Jalan Ayer Keroh Lama 75450, Melaka , Malaysia
E-mail: subar.kannan@mmu.edu.my

[3] Electrical Electronic Engineering Department
University Technology Petronas Bandar Seri Iskandar, Tronoh, Perak , Malaysia
E-mail: vijanth_sagayan@petronas.com.my

ABSTRACT

Data warehouse is a database with tools that stores current and historical data of potential interest. One of them which, will be investigated in this paper, is the state school data. Mining or analyzing data extract from such system would be interesting for possible smart school application. This paper looks at the framework of virtual smart school implementation based on students' exam results.

Keywords

Data Warehousing, Data Mining, Online Analytical Processing (OLAP)

1. INTRODUCTION

Currently it has been a trend that most companies build enterprise-wide data warehouses where a central data warehouse serves the entire organization or they can create smaller decentralized warehouses called data marts (Kimball, 1996). A data mart is a subset of a data warehouse in which a summarized or highly focused portion of the organization's data is placed in a separate database for users with specific interest (Roussopoulos et al, 1995). The paper describe the implementation of the data mining and warehousing system for the application for possible smart school project using state school data as illustrated in Figure 1.

The paper is organized as follows. The second section of the paper give an overview of data warehousing and some phrases or terms used throughout the paper. The third section will discuss on the OLAP (online analytical processing) as data mining tool for retrieving information from data warehouse. The drill-down approach for multidimensional OLAP (MOLAP) is discussed in brief in the fourth section and finally section five concludes the paper.

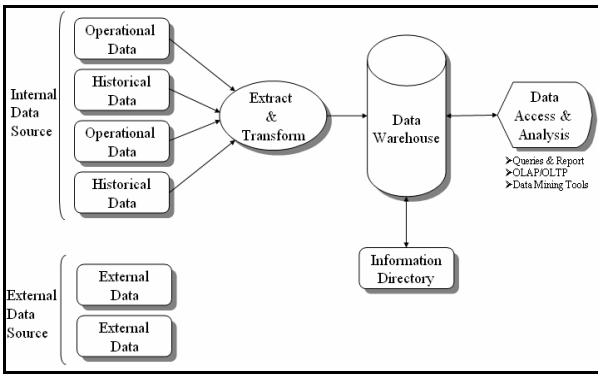

Figure 1 Component of Data Warehouse

2. DATA WAREHOUSING

Data Warehouse is a type of database system aimed at effective integration of operational databases into an environment that enables strategic use of data (Zhou et al, 1995). These technologies include relational and multidimensional database management systems, client/server architecture, meta-data modeling and repositories, graphical user interface and much more (Hammer et al, 1995),(Harinarayan et al., 1996). This informational database system which is currently much subject of researched is not only commonly used in business or finance sector but can be applied appropriately in educational sectors too by mining large pool of data for wise decision making as shown in Figure 2. Information normally is crucial and analyzing or processing this information using some tools is essential for critical decision-making

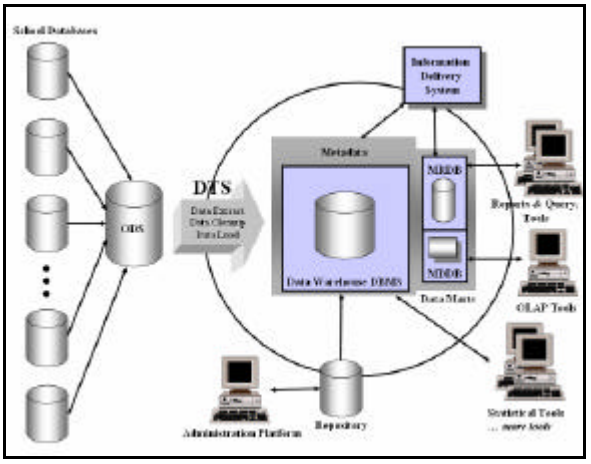

Figure 2 Data Warehouse Architecture

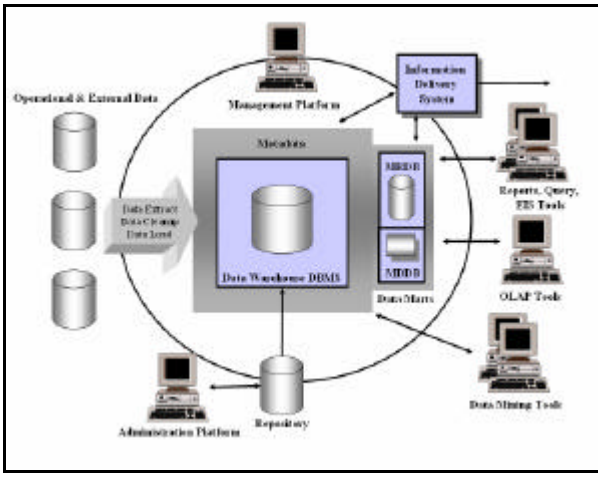

Figure 3 Analyzing components extracting Data Warehouse

3. POPULATING THE DATABASE WITH USEFUL INFORMATION

Typically, a data warehouse is populated with historical information from within a particular organization (Bunger et al., 2001). This need not be followed strictly and, in many circumstances, a company's warehouse tables may be populated with a wide variety of data sources, often including data providing information concerning a competitor business. Collecting all of this different data and storing it in one place is an extremely challenging process (Calvanese et al., 2001).

Data Transformation Services (DTS) assist in the import and export of data between heterogeneous data sources, using OLE-DB (object linking and embedding database) - based architecture (Zhou et al., 1995). Thus, transformation is used to populate a warehouse and to update the data in the warehouse. However, there is no lower restriction on the size of the data warehouse.

4. RETRIEVING INFORMATION FROM DATA WAREHOUSE

The process of retrieving data from the warehouse can vary greatly depending on the desired results. There are many form of possible retrieval from a data warehouses and it is flexibility that will drive how this retrieving process can be implemented. There are many tools for retrieving the data warehouse, such as building simple query and reporting through SQL (special query language) statements and using Crystal Reports, Microsoft Visual Basic, Report Generator or some third party tool as shown in Figure 3. The tools may expand to OLAP and data mining, where the structure includes many more third party tools. There are many inherent problems associated with data, which includes the limited amount of portability, and the often-vast amount of data that must be sifted through for each query (Nicola & Jarke, 2000).

For the state school system uses to transport all the 14 different schools databases into the data warehouse, (which represents fourteen different states in Malaysia), it will first load to the Operational Data Store (ODS), as a backup by using the Data Transformation Services (DTS). The data warehouse architecture shown in clearly identifies eight important data warehouse components that every data warehouse should have, which includes:

- Data sourcing, cleanup, transformation and migration tool using DTS concepts
- Meta-data repository
- Warehouse database technology
- Data marts
- OLAP technology
- Data query, reporting, analysis and statistical tools
- Data warehouse administration
- Information delivery system

Before the data can be loaded into the data warehouse, it must be first be transformed into an integrated, consistent format (Nicola & Jarke, 2000). A transformation is the sequence of procedural operations that are applied to the information in a data source before it can be stored in the specified destination. In this system, Data Transformation Services (DTS) is used to transports data because it can supports many types of database such as text files, excel spreadsheets, access and network databases. For the smart school database system, please refers to Figure 4 for the DTS mapping from the 14 schools databases into the data warehouse.

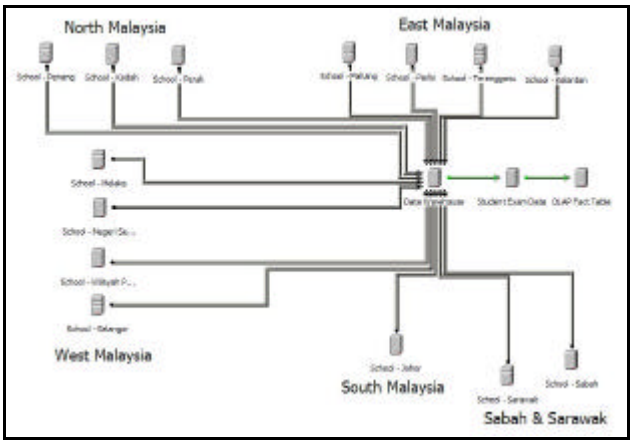

Figure 4 Components of Data Warehouse

5. OLAP FOR SMART SCHOOL

During the last ten years, a significant percentage of corporate data has migrated to relational database. Relational databases, have been used heavily in the areas of operations and control, with a particular emphasis on transaction processing (for example, manufacturing process control, brokerage trading).

It is important to distinguish the capabilities of a data warehouse from those of an OLAP (On-line Analytical Processing) system (Shoshani, 1997). In contrast to a data warehouse, which is usually based on relational technology, OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for further analysis.

OLAP enables analysts, managers, and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information. OLAP transforms raw data so that it reflects the real dimensionality of the enterprise as understood by the user.

The major convenience of OLAP tools is their ability to dynamically represent the data without re-querying the database (Nicola & Jarke, 2000). For example, the data summarized by year, if a particular year summary looks peculiar, then the details for that specific year can be viewed with a click of a button, while the remaining years remain in summary view.

There is one type of OLAP, known as MOLAP, which specifications are based on the multiple degree of normalization in the database tables (Harinarayan et al., 1996).

6. MULTIDIMENSIONAL OLAP

MOLAP partitions store aggregations and a copy of the source data (fact and dimension data) in a multidimensional structure on the analysis server. All partitions are stored on the analysis server. Analysis Services responds to queries faster with MOLAP than with any other storage mode for the following reasons:

Compression: Analysis Services compresses the source data and its aggregations to approximately 20 percent of the size of the same data stored in a relational database. The actual compression ratio varies based on a variety of factors, such as the number of duplicate keys and bit encoding algorithms. This reduction in storage size enables analysis services to resolve a query against fact-level data or aggregations stored in a MOLAP structure much faster than against data and aggregations stored in a relational structure because the size of the physical data being retrieved from the hard disk is smaller.

Multidimensional data structures: Analysis Services uses native multidimensional data structures to quickly find the fact data, either at the fact level or at higher aggregation levels, which is depicted in Figure 5.

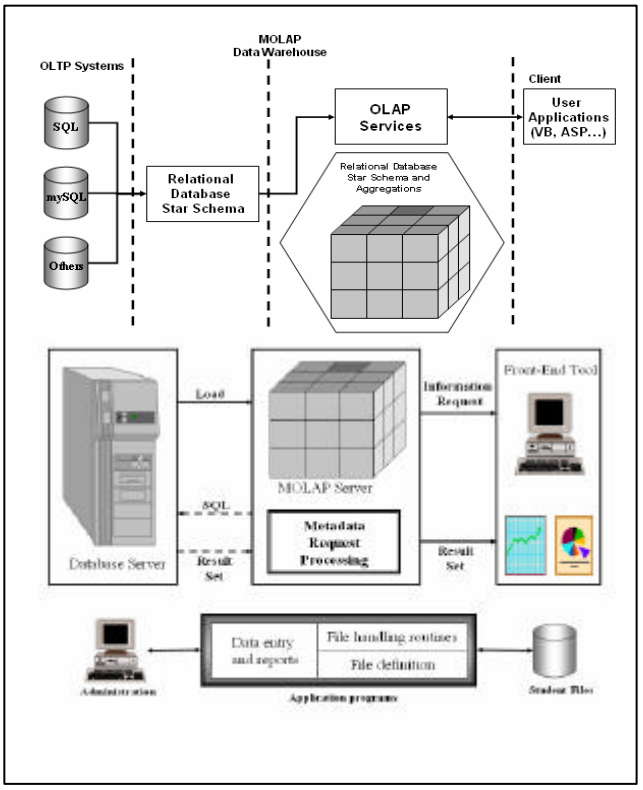

Figure 5 MOLAP Architecture

Data in a single service: MOLAP partitions are generally stored on a single Analysis server, with the relational database frequently stored on a server separate from the Analysis server. Analysis Services must query across the network whenever it needs to access the relational tables to resolve a query. The impact of querying across the network depends on the performance characteristics of the network itself. Even when the relational database is placed on the same server as Analysis Services, inter-process calls and the associated context switching are required to retrieve relational data. With a MOLAP partition, calls to the relational database, whether local or over the network, do not occur during querying.

7. OLAP METHODOLOGY

The foundation of OLAP is the ability to "drill down" into the data, as far into its minute detail as is necessary to get the answers you need. With OLAP, we can look beyond the summary data for details of what may have caused the trend demonstrated by the summary. While there are tools available to assist in finding trends and the reason for trends and the reasons for the reasons – a human component is still required. Someone has to make sense of the data and this can take hours of close observations. Below are a few terms about the OLAP concepts:

Cubes: Modeling data multi-dimensionally is a way to facilitate online business analysis and query performance (V.Harinarayan et al., 1996). The OLAP Manager allows you to turn data stored in relational databases into meaningful, easy to navigate business information by creating a data cube. Cube concepts and terminology are described in the following.

Relational Schemas and Cubes: The most common way of managing relational data for multidimensional use is with a star schema. A star schema consists of a single fact table that is joined to a number of dimension tables. The fact table contains the numeric data that corresponds to the measures of a cube. Dimension table columns, as their name implies, map to the hierarchical levels in a dimension. Note: A star schema is not required in order to create a cube (V.Harinarayan et al., 1996). You can also use a snowflake schema, or even a single table schema.

Dimensions of a Cube: The dimensions of a cube represent distinct categories for analyzing business data. Categories such as time, geography, or product line breakdowns are typical cube dimensions.

Dimensions and Hierarchies: Dimensions are typically organized into hierarchies of information that map to columns in a relational database (Bunger et al., 2001). Dimension hierarchies are grouped into levels consisting of dimension members. Each level in a dimension can be rolled together to form the values for the next highest level. For example, in a time dimension, days roll into months, and months roll into quarters.

Measures of a Cube: Measures are the quantitative values in the database that you want to analyze. Typical measures are sales, cost and budget data. Measures are analyzed against the different dimension categories of a cube. For example, you may want to analyze sales and budget data (your measures) for a particular product (a dimension) across various countries (specific levels of a geography dimension) during two particular years (levels of a time dimension).

8. MOLAP SERVER (DRILL-DOWN APPROACH)

MOLAP requires the data to be stored in a multidimensional format. It involves the creation of multidimensional blocks called data cubes (Harinarayan et al., 1996). The cube in Figure 5 MOLAP Architecture may have three axes (dimensions), or it may have more. Each axis (dimension) represents a logical category of data. One axis may for example represent the geographic location of the data, while others may indicate a state of time or a specific school. Each of the categories, which will be described in the following section, can be broken down into successive levels and it is possible to drill up or down between the levels.

Geographical Dimension

If a school is located in Malaysia, this dimension may have the following levels:

- Malaysia, which can be broken down into:
- Regions, which can be broken down into:
- States, which can be broken down into:
- Cities

School Dimension

If the school has lesson taught and classrooms, this dimension may contain:

- School, which can be broken down into:
- Classroom, which can be broken down into:
- Subject, which can be broken down into:
- Result

Time Dimension

If the school follows a fiscal year, we have the following levels:

- Fiscal year, which can be broken down into:
- Quarters, which can be broken down into:
- Months, which can be broken down into:
- Weeks

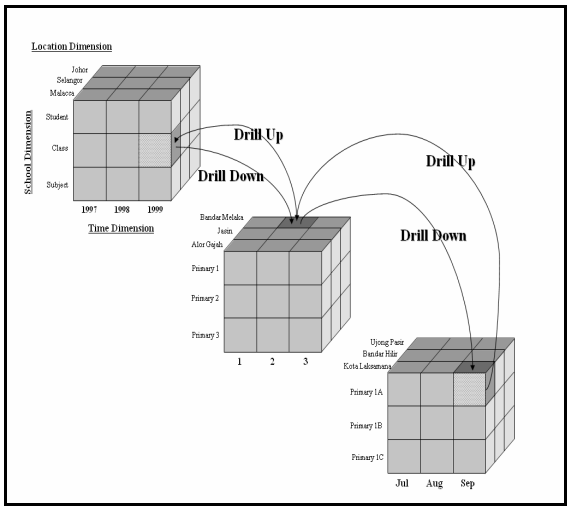

Figure 6 OLAP Drill-Down/Drill-Up Concepts

9. CONCLUSIONS

This paper presents a conceptual framework in data warehousing and data mining approaches on smart school prototypes system. The paper looks into how data from various locations being transferred into large data warehouse before the data being analyzed. The paper also described the methodology and drilling up/down processes for OLAP (online analytical processing) and its multidimensional form known as MOLAP for smart school system.

REFERENCES

Bunger, C. J., Colby, L. S., Cole, R. L., McKenna, W. J., Mulagund, D. G., & Wilhite. (2001). *Aggregate maintenance for data warehousing in Informamix Red Brick Vista*. Paper presented at the Proceedings of the 27th VLDB conference.

Calvanese, D., Giacomo, D., Lenzerini, G., Nardi, D., & Rosati, R. (2001). Data integration in data ware housing warehousing. *International Journal of Cooperative Information Systems*.

Zhou, G., Hull, R., King, R., & Franchitti. (1995). Supporting Data Integration and Warehousing Using H2O. *IEEE Data Engineering Bulletin, 18*(2), 29-40.

Hammer, J., Garcia-Molina, G., Labio, W., Widom, J. & Zhuge, Y. (1995). The Stanford Data Warehousing Project. *IEEE Data Engineering Bulletin, 18*(2), 41-48.

Kimball, R. (1996). *The Data Warehouse Toolkit*: John Wiley.

Roussopoulos, N., Chen, C. M., Kelley, S., Delis, A., & Papakonstantinou, Y. (1995). The Maryland ADMS Project: Views R Us. *IEEE Data Engineering Bulletin, 18*(2), 19-28.

Nicola, M., & Jarke, M. (2000). Performance Modelling of distributed and replicated databases Research Survey. *IEEE Transactions Knowledge and Data Engineering, 12*(4), 214-233.

Shoshani, A. (1997). OLAP and Statistical Databases: Similarities and Differences. *Proc. ACM PODS*, 185-196.

Harinarayan, V., Rajaraman, A., & Ullman, J.D (1996). Implementing Data Cubes Efficiently. *Proc. ACM SIGMOD*, 205-216.