Data Warehousing and Data Mining

Ms. Jeevantika Lingalwar[1], Ms Kajal Choudhary[2], Ms. Palak Agrawal[3], Ms. Aboli Patil[4]
S.B.J.I.T.M.R

***Abstract*—*Data-driven* decision support systems, such as data warehouses can serve the requirement of extraction of information from more than one subject area. Data warehouses standardize the data across the organization so as to have a single view of information. Data warehouses can provide the information required by the decision makers. Developing a data warehouse for educational institute is the less focused area since educational institutes are non-profit and service oriented organizations. In present day scenario where education has been privatized and cut throat competition is prevailing, institutes needs to be more organized and need to take better decisions. Institute's enrollments are increasing as a result of increase in the number of branches and intake. Now a day, any reputed Institute's enrollments count in to thousands. In view of these factors the challenges for the management are meeting the diverse needs of students and facing increased complexity in academic processes. The complexity of these challenges requires continual improvements in operational strategies based on accurate, timely and consistent information. The cost of building a data warehouse is expensive for any educational institution as it requires data warehouse tools for building data warehouse and extracting data using data mining tools from data warehouse. The present study provides an option to build data warehouse and extract useful information using data warehousing and data mining open source tools. In this paper we have explored the need of data warehouse / business intelligence for an educational institute, the operational data of an educational institution has been used for experimentation. The study may help decision makers of educational institutes across the globe for better decisions..**

***Index Terms*—*Data* warehouse, data mining, analysis, ETL, BI.**

I. Introduction

Now a day, the educational institutes have to generate funds for their research and other operational activities as the government funding has been limited to aided institutes. Utilizing a decision support system is a proactive way to use data to manage, operate, and evaluate educational institute in a better way. Depending on the quality and availability of the underlying data, such a system could address a wide range of problems by distilling data from any combination of education records maintenance system. The data mining from data warehouse can be a ready and effective system for the decision makers. A data warehouse is a subject oriented integrated, non-volatile, and time variant collection of data in support of management decisions [1].Data warehouse obtains the data from a number of operational data base systems which can be based on RDBMS/ERP package, etc. The data from these sources are converted into a form suitable for data warehouse. This process is called Extraction, Transformation and Loading (ETL). In addition to the target database, there will be another database to store the metadata, called the metadata repository. This data base contains data about data-description of source data, target data and how the source data has been modified into target data. The client software will be used to generate reports.

II. Motivation

Motivation for building data warehouse for the educational institute is from two sources, internal sources like inability of current operational systems to provide required information for parameter driven analysis and external sources like competitive factors. A survey is carried out by visiting several educational institutes to gather information regarding the current practices the institutes have implemented as decision support systems. The findings are summarized below.

1) The data is stored in different sources in distributed locations.
2) Users find difficulty in locating the reports needed by them.
3) The user interface for the current operational system is not satisfactory and is confusing and hard to use for decision makers.
4) When the consolidated report from two or more different subject area is required, it is almost impossible.
5) There is no easy way to get assistance.

Utilizing a decision support system is a proactive way to use data to manage, operate, and evaluate educational institute in a better way. Depending on the quality and availability of the underlying data, such a system could address a wide range of problems by distilling data from any combination of education records maintenance system. The purpose of this paper is to investigate current system of information delivery and propose a better system for timely, accurate, consistent information delivery to the decision makers of the educational institutes .

III. Literature Review

Following section briefly describes the different application areas for which data warehouses are built.

A. Retail Sales

Data is collected at several interesting places in a grocery

store. Some of the most useful data is collected at the cash registers as customers purchase products. Modern grocery store scans the bar codes directly into the point_of_sale system. The POS system is at the front door of the grocery store where consumer takeaway is measured. The back door, where vendors make deliveries, is another interesting data collection point [8].At the grocery store, management is concerned with logistics of ordering, stocking, and selling products while maximizing profit. Some of the most significant management decisions are on pricing and promotions. Both store management and marketing spend a great deal of time tinkering with pricing and promotions. In such scenarios, data warehouses come to rescue.

B. Telecommunications

A telecommunications company generates hundreds of millions of call-detail transactions in a year. For promoting proper products and services, the company needs to analyze these detailed transactions. The data warehouse for the company has to store data at the lowest level of detail.

C. Transportation

In this case, the airline's marketing department wants to analyze the flight activity of each member of its frequent flyer program. The department is interested in seeing which flights the company's frequent flyers take, which planes they fly, what fare basis they pay, how often they upgrade, how they earn. These requirements can be fulfilled by data warehouse.

D. Education

There are some efforts in the area of data warehouse for building data warehouse for education domain. The paper by Carlo DELL'AQUILA [10] summarizes the experience in designing and modeling an academic data warehouse. Existing facilities and databases affect the chosen data warehouse that brings them together to support decisional activities leading the whole university environment, including administrators, faculties and students. The choice to develop a dedicated system is mainly forced by the peculiar information type that defines the basic information in data warehouse widely different from institution to institution .In the article titled 'What academia can gain from building a data warehouse' by David Wierschem, et.al [11].The authors have identified the opportunities associated with developing a data warehouse in an academic environment. They begin by explaining what a data warehouse is and what its informational contents may include, relative to the academic environment. Next they addressed the current environment drivers that provide the opportunities for taking advantage of a data warehouse and some of the obstacles inhibiting the development of an academic data warehouse. Finally, the article provides strategies to justify developing a data warehouse for an academic institution.

IV. DATAWAREHOUSE ENVIRONMENT

Utilizing a decision support system is a proactive way to use data to manage, operate, and evaluate educational institute in a better way. Depending on the quality and availability of the underlying data, such a system could address a wide range of problems by distilling data from any combination of education records maintenance system. The data mining from data warehouse can be a ready and effective system for the decision makers. The data from reputed engineering college namely R V College of Engineering, Bangalore, Karnataka, India, has been considered for this study. Fig. 1 shows the data warehouse architecture of RV College where source systems are smart campus, asset management server and csv files, the information is spread across diverse platforms, data from different sources is collected and then consolidated to produce required report. ETL activities are performed to extract the data from heterogeneous sources and load into staging and then load the data into dimension and fact tables as per the schedules. We proceed to extract the BI report from data warehouse on demand based on requirement from the management. In an educational institute, main information required will be regarding key components of the education institute, namely students, employees and infrastructure. The purpose of this paper was to investigate current system of information delivery and proposing a better system for timely, accurate, consistent information delivery to the decision makers of the educational institute. The paper has been prepared in order to extend the usage of current available technology in decision making processes of educational institute.

Fig. 1. Engg_Data warehouse architecture

Data warehouse enables the decision makers with benefits listed below.

1) Phenomenal improvements in turnaround time for data access and reporting
2) Standardizing data across the organization so that there will be one view of information.
3) Merging data from various source systems to create a more comprehensive information source.
4) Reduction in costs to create and distribute information and reports.
5) Encouraging and improving fact-based decision making.

V. BI-REPORTING

This refers to the variety of capabilities that can be provided to the users to leverage the presentation area for analytic decision making. All data access tools query the data in the data warehouse presentation area. A data access tool can be as simple as an ad hoc query tool or as complex as sophisticated data mining application. The majority of the users use pre-built parameter driven analytic applications to

access the data. This enables them to retrieve the required information and analyze hidden pattern in the retrieved data [11].Using suitable data mining techniques, the useful information can be extracted from the data warehouse. Data mining form three main components of the institute, namely Employees, Students and Infrastructure. Employee data mart can provide the users with the information such as career growth and attrition rate. Student mart can provide the information related to the student like best outgoing student considering his academic and non academic activities. Information regarding assets such as the investment in a particular financial year can also be accessed.

VI. Results

Once the data warehouse is deployed, it invariably becomes a mission-critical application. Users depend on the data warehouse to provide them with the information they need to function properly. To make certain that the ETL process runs and completes, it must be actively monitored and supported. Some of the results observed after querying the data marts are documented below. The results are cross checked with the requirements specified by the different types of users. The requirements with regard to asset information were to extract the information regarding the number of assets of each type in the Institute. The different data marts are queried using SQL query. The results returned by the queries are found accurate and meeting users demands. The sample screen shots of queries and the results are shown.

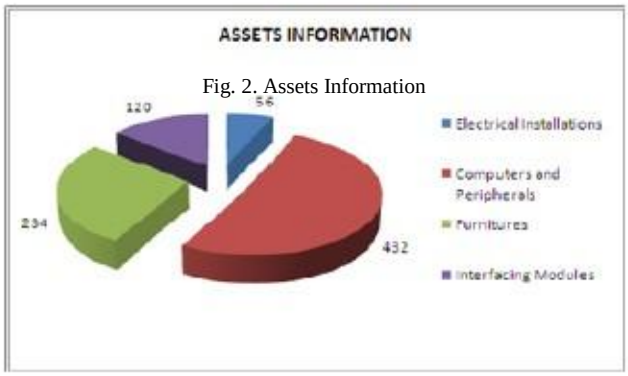

Fig. 2. Assets Information

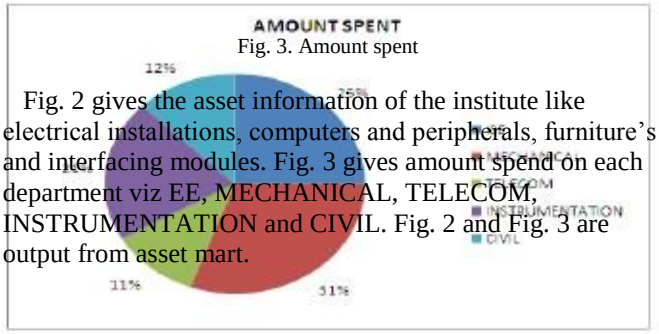

Fig. 3. Amount spent

Fig. 2 gives the asset information of the institute like electrical installations, computers and peripherals, furniture's and interfacing modules. Fig. 3 gives amount spend on each department viz EE, MECHANICAL, TELECOM, INSTRUMENTATION and CIVIL. Fig. 2 and Fig. 3 are output from asset mart.

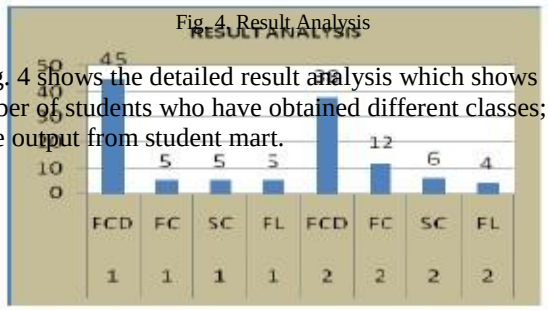

Fig. 4. Result Analysis

Fig. 4 shows the detailed result analysis which shows number of students who have obtained different classes; this is the output from student mart.

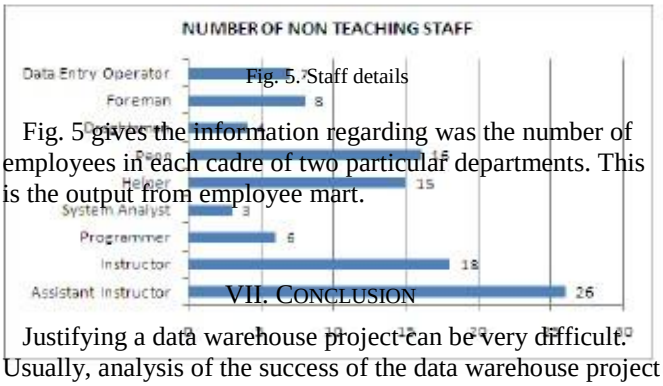

Fig. 5. Staff details

Fig. 5 gives the information regarding was the number of employees in each cadre of two particular departments. This is the output from employee mart.

VII. Conclusion

Justifying a data warehouse project can be very difficult. Usually, analysis of the success of the data warehouse project is done considering the financial benefits against the investment. Since most of the educational institutes are nonprofit organizations and service oriented, the evaluation of the usefulness of the data warehouse can be done on the basis of its ability to meet user's requirements. The academic data which was spread all across different sources has been loaded into single platform. The decision makers can extract information regarding three main components of the institute, namely Employees, Students and Infrastructure. Employee data mart can provide the users with the information such as career growth and attrition rate. Student mart can provide the information related to the student like best outgoing student considering his academic and non academic activities. Information regarding assets such as the investment in a particular financial year can also be accessed. In educational institute, decision makers ask "What are the expected results and benefits?" when making a data warehouse project rather than "What is the anticipated return on investment?". The data warehouse developed has met their expectations. Benefits of the present project can be more if the Institute has positive approach towards new technologies. They can take micro-level decisions in a timely manner without the need to depend on their IT staff. They can perform extensive analysis of stored data to provide answers to the exhaustive queries to the administration cadre. This helps them to formulate

strategies and policies for employees and students. This helps students and Employees in making decisions. They are the ultimate beneficiaries of the new policies formulated by the decision makers and policy planner's extensive analysis on student and employee related data.
Over all 80 to 85% of decisions are made based on the reports generated by the proposed system. The realistic productivity is about 85%.

VIII. Future Scope

The enhancement that can be carried out on the present system is the implementation of the real time ETL system. Real time ETL refers to the software that moves data synchronously into a data warehouse with some urgency-within minutes of the execution of the business transaction. Implementation of real-time data warehouse reflects a new generation of hardware, software and techniques. Capture, Transform, and Flow (CTF) is a relatively new category of data integration tools designed to simplify the movement of real-time data across heterogeneous database technologies. The transformation functionality of CTF tools is typically basic in comparison with today's mature ETL tools, so often real time data warehouse CTF solutions involve moving data from the operational environment, lightly transforming it using the CTF tool and then staging it.

Acknowledgment

Authors would like to acknowledge and extend our heartfelt gratitude to the following persons who have made the completion of this research paper possible: Technical Staff of R V College of Engineering, for their vital support.Sri B Sridhara Murthy, Business Intelligence Analyst, HP, Bangalore for the much needed motivation. Sri B M Sagar, Assistant Professor, R V College of Engineering, Bangalore, India, Mr.Shahzad, SME, CSC USA,Mr. Parswanath Project Manager(Data Warehousing Wing). Wipro Technologies, India. Mrs. Radha Sarvana, Analyst, Wipro Technologies, India.

References

[1] Ralph Kimball, the Data Warehouse ETL Toolkit, *Wiley India Pvt Ltd.,* 2006.
[2] KV. K. K. Prasad, Data warehouse development Tools,*Dreamtech Press,* 2006.
[3] W. H. Inmon, Building the Data Warehouse. *Wiley; 3rd edition* March 15, 2002.
[4] Alex Berson, Data Warehousing Data Mining & OLAP, Computing *Mcgraw-Hill,* November 5, 1997.
[5] Arshad Khan, SAP and BW Data Warehousing, *Khan Consulting and Publishing, LLC* (January 1, 2005)
[6] Carlo DELL'AQUILA,'An Academic Data Warehouse' *World Scientific and Engineering Academy and Society (WSEAS) Stevens Point,* Wisconsin, USA ©2007.
[7] McMillen and Randy McBroom.,'what academia can gain from building a data warehouse' no.1,pp.41-46.2008
[8] Channah F. Naiman, Aris M. Ouksel "A Classification of Semantic Conflicts in Heterogeneous Database Systems", *Journal of Organizational Computing,* vol. 5, 1995.
[9] John Hess, "Dealing With Missing Values In The Data Warehouse" *A Report of Stonebridge Technologies, Inc-1998.*
[10] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."Analysis of Data Quality Aspects in DataWarehouse Systems",(IJCSIT) *International Journal of Computer Science and InformationTechnologies,* vol. 2 (1) , 2010, 477-485.
[11] Jaideep Srivastava, Ping-Yao Chen,"Warehouse Creation-A Potential Roadblock to Data Warehousing", *IEEE Transactions on Knowledge and Data Engineering* January/February 1999 (vol. 11, no. 1) pp.118-126.
[12] Amit Rudra,Emilie Yeo (1999) "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd *HawaiiInternational Conference on System Sciences* – 1999.
[13] Jesús Bisbal et all,"Legacy Information Systems: Issues and Directions", *IEEE Software* September/ October 1999.
[14] Scott W. Ambler "Challenges with legacy data: Knowing your data enemy is the first step in overcoming it", *Practice Leader-2001.*