Enhancing Student Data Warehouse by Implementing Dataset Level Transformation in ETL Process

A Nwe Soe, Khin Mar Myo
University of Computer Studies (Mawlamyine)
ans.anwe@gmail.com, kmmyo09@gmail.com

Abstract

Today, every organization is needed not only to record but also to manage and analyze a huge amount of data. Data warehouse and OLAP techniques are valuable tools for today's competitive, fast evolving world. There have been done many researches to improve the effectiveness of data warehouse. While some works are focused on back-end tools by enhancing the ETL processes, others emphasis on front-end tools by enhancing the OLAP processes. This paper tries to understand on Data warehouse and OLAP technology by implementing the Student Data warehouse for all Universities of Computer Studies, Myanmar. This paper focuses on the data preprocessing step by conducting the dataset level transformation in ETL portion. Finally, this paper describes the analysis reports by using OLAP processes.

1. Introduction

In current era, both commercial and scientific applications are working out on the huge amount of data information. And, they need to extract some decision reports based on data warehouse. A data warehouse is a copy of transaction data specifically structured for querying and reporting. A data warehouse can be normalized or denormalized. It can be a relational database, multidimensional database, flat file, hierarchical database, object database, etc. Data warehouse data often gets changed. And data warehouses often focus on a specific activity or entity. Data warehousing is not necessarily for the needs of "decision makers" or used in the process of decision making.

In general, the queries information can be extracted from data warehouse by using OLAP operations of DBMS. OLAP operations are classified into a class of complex queries in which group-by and aggregate operations are held on the star schema, and they are characterized by the analysis of merged enterprise data supporting and user's analytical and navigational activities. They can support the data analysis and data summarization from multiple attribute values of database tables.

Data preprocessing task is very important not only for data mining but also for data warehouse. Because the data getting from different sources may contain missing values, duplicate records, noisy data, etc. There have been developed many methods and techniques for data preprocessing, such as data cleaning, data integration, data transformation and data reduction.

According to [2], there are three transformation levels in ETL. They are row level transformation, dataset level transformation and data warehouse level transformation.

This paper describes how to conduct the data transformation for enhancing the ETL process in student data warehouse creation. Data transformation is conducted at data set level in this system. This paper also describes the detail of step by step construction of data warehouse. Firstly data from transaction database are modeled into multiple dimensions and schema. Then, data are physically stored in relational OLAP server (ROLAP).

The rest of the paper is organized as follows: related work is described in section 2. Section 3 describes background theory and Section 4 presents design for Student Data Warehouse. System Implementation comes in Section 5 with transformation, modeling and schema. Section 6 will describe ROLAP with sample queries and section 7 describes the conclusion.

2. Related Work

According to [5], this paper describes creating data warehouse for clinical information system. In their paper, Data Warehouse was structured using star schema for building dataset of clinical information system. Users of clinical information system want to analyze the data summarized to various level using OLAP processes.

Recently public administration has become aware of the benefits of data warehousing [4]. The new technology enables interactive data analysis and ad hoc reporting. One area of interest for public administration is to control the higher educational system. This system has a lot of characteristics in common with a distributed organization. Furthermore, each university acts widely autonomously and competes with other universities for students. In this context, a data warehouse can help to provide a fair allocation of available funds or personal resources to all universities. A detailed analysis of the number of students at each university can be accomplished by a data warehouse system.

Data warehouse contains vast amount of information. Most of the data warehouses are concerned with business's important data. The paper [1] pointed out the important of security factor in data warehouse. Therefore security should be brought to the attention of both data warehouse administrators and managers who decide on who should have access to what data. Many data warehouses are implemented on relational databases the already offer a wide variety of security mechanisms such as authentication, access control and auditing.

3. Background Theory

This section describes some background theories concerning with other work. They are data warehouse architecture, ETL process, staging area and design steps for data warehouse creation.

3.1 Data Warehouse Architecture

The author [3] stated data warehouse architecture with three tiers. The author also stated another architectures for data warehouse: (i) basic architecture, (ii) architecture with staging area and (iii) architecture with staging and data mart.

The following Figure 1 illustrates the data warehouse architecture with staging area. This architecture uses the staging area to conduct the ETL process.

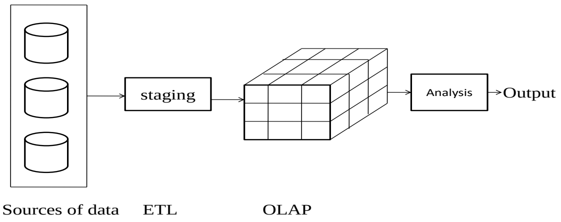


Figure 1. Data Warehouse Architecture with Staging

3.2 ETL process

Data warehouse system uses back-end tools and utilities to populate and refresh their data. These tools and utilities include data extraction, data cleaning, data transformation, loading and refresh. These all supporting processes to back-end tools are called ETL processes.

The first part of ETL process is to *extract* the data from various source systems. Data source format can be relational database, flat file and, non-relational database. The *transformation phase* applies a number rules to be extracted so as to convert different data format into single format. The *loading phase* loads the transformed data into data warehouse. So that it can be use for various analytical purposes [3].

3.3 Staging Area

The staging area is a temporary area for conducting ETL process. According to [2], ETL staging environment is the factory in which data is captured and transformed into what will become a data warehouse. By incorporating ETL staging principle in the extract, transform and load applications and ETL applications can know what is in the assembly line now, where and how it came from, when and how it is going, and what to do about it.

3.4 Steps for Data Warehousing

According to [6], the logical design and physical design of data warehouse must be considered. Multidimensional data model is a typically use for the design of not only corporate data warehouse but also data mart. Multidimensional model allows to highlight facts (measures) which represent useful indicators for the administrators.

Such a model can adopt star schema, snow flake schema or fact constellation schema. After the logically designing, these models must be considered where and how the data are physically stored. Mostly three types of OLAP servers can be used to store these data: (i) relational OLAP (ROLAP), (ii) multidimensional OLAP (MOLAP) and (iii) hybrid OLAP (HOLAP).

4. Design for Student Data Warehouse

This section describes the detail of system design which is based on warehouse architecture with staging area as described in section 3. The main idea of our system architecture is to use the staging area for data transformation. There are 3 main phases in system design which is illustrated in Figure 2.

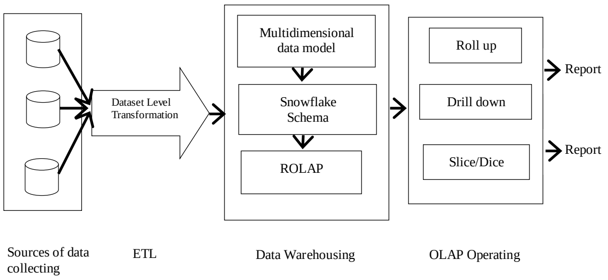


Figure 2. System Design

4.1 Data Transformation

The main responsibilities of this component are to extract the transactional data from different sources and load onto staging area for data transformation. In this phase, dataset level transformation is performed.

4.2 Data Warehousing

There are three main tasks in this phase. Firstly, data must be modeled as multi-dimension. Secondly, database schema must be considered for data. Finally, the data must be recorded in Relational database (ROLAP).

4.3 Data Analyzing

Data from multi-dimensional model (cube) are analyzed in this phase by conducting the basic operations of OLAP. Such as drill down, roll up, slice and dice operations.

5. System Implementation

Student data warehouse system is implemented to evaluate the effectiveness of the system design from section 4. This section describes the detail of the system implementation.

5.1 Source of Data

The student data from the whole Universities of Computer Studies, Myanmar are chosen as source of data for Warehouse creation and analyzing. There are totally about 26 Computer Universities in Myanmar under the Ministry of Science and Technology. Although there may have to store many range of data. This system only emphasizes on student data. Example tables for students are as follow:

(1)Student (No, Rno, Name, Address, D-O-B, Parent, Mark)
(2)Degree (No, Rno, Name, Degree, CU, Academicyear)

Table 1. Student Table

No	Rno	Name	Address	D-O-B	Parent	Mark
1	2cs1	Ma Su Mon	Yangon	03:20:1987	U Mya	568
2	2cs2	Ma Yu Mon	Yangon	05:24:1987	U Hla	560
3	2cs3	Ma Thuzar	Pathein	04:03:1986	U Thet	540
4	2ct1	Ma Htet	Maupin	06:06:1985	U Htay	530
5	2ct2	Ma Cherry	Hpaan	07:23:1984	U Than	500
6	2ct3	Ma Nandar	Dawei	06:29:1986	U Linn	590
7	3cs1	Mg Kyaw	Myeik	06:30:1987	U Zaw	570
8	3cs2	Mg Zaw Lin	Sittway	07:29:1986	U Mya	560
9	3cs3	Ma Yamin	Pinlon	05:28:1986	U Thein	540
10	3cs4	Mg Min Thu	Lashio	05:22:1987	U Win	545
11	3ct1	Ma Aye Thu	Loikaw	06:19:1986	U Kyaw	550
12	3ct2	Ma Hnin Yu	Kalay	05:23:1987	U Min	540
13	3ct3	Ma Sabai	Bamaw	06:09:1986	U Aung	548
14	3ct4	Ma Thazin	Hinthada	12:04:1984	U Khine	540

Table 2. Degree Table

No	Rno	Name	Degree	CU	Result	AY
1	3cs1	Ma Theingyi	B.C.Sc	Yangon	Pass	2007
2	3cs2	Ma Ei Mon	B.C.Sc	Kalay	Pass	2007
3	3cs3	Ma Thet Su	-	Myike	Fail	2008
4	3ct1	Ma Soe San	B.C.Tech	Yangon	Pass	2008
5	3ct2	Ma Cherry	B.C.Tech	Dawei	Pass	2008
6	3ct3	Ma Wittyi	-	Lashio	Fail	2000
7	4cs1	Mg Lin Lin	B.C.Sc(Q)	Yangon	Pass	1998
8	4cs2	Mg Zaw Htin	B.C.Sc(Q)	Yangon	Pass	1998
9	4cs3	Ma Yamin	B.C.Sc(Q)	Pathein	Pass	2000
10	4cs4	Mg Min Thu	B.C.Sc(Q)	Yangon	Pass	2001
11	DS1	Ma Thu Thu	D.C.Sc	Bamaw	Pass	2001
12	DS2	Ma Hnin Yu	-	Yangon	Fail	2002
13	DS3	Ma Phyo Oo	D.C.Sc	Kalay	Pass	2003
14	DS4	Ma Marlar	-	Yangon	Fail	2003

5.2 Data Transformation

Dataset getting from different Universities contains not only interested attributes but also uninterested attributes. This phase is responsible to transform the dataset into necessary format.

Firstly, dataset are loaded onto staging area then perform two main tasks. The first one is to make the same format of D-O-B field from dataset. The value of D-O-B field may have different formats such as (12:03:2009), (December 3 2009), etc. In Table 1, the values of shaded attributes (D-O-B) are transformed into same format. Some of the rules for transformation are:

(i) If (day, month, year) then (month, day, ,year)
(ii) If (month(text), day, year) then (month, day, year)

After this transformation, table 1 is changed as illustrated in table 3.

In table 2, the aggregation function is conducted on these transactional databases. In table 2, there are totally 14 records. Dataset is aggregated based on Degree field. This table 2 becomes as illustrated in table 4.

Table 3. Student Table

No	Rno	Name	Address	Parent	D-O-B	Mark
1	2cs1	Ma Su Mon	Yangon	U Mya	March 20 1987	568
2	2cs2	Ma Yu Mon	Yangon	U Hla	May 24 1987	560
3	2cs3	Ma Thuzar	Pathein	U Thet	April 03 1986	540
4	2ct1	Ma Htet Htet	Maupin	U Htay	June 06 1985	530
5	2ct2	Ma Cherry	Hpaan	U Than	July 23 1984	500
6	2ct3	Ma Nandar	Dawei	U Linn	June 29 1986	590
7	3cs1	Mg Kyaw	Myeik	U Zaw	June 30 1987	570
8	3cs2	Mg Zaw Lin	Sittway	U Mya	July 29 1986	560
9	3cs3	Ma Yamin	Pinlon	U Thein	May 28 1986	540
10	3cs4	Mg Min Thu	Lashio	U Win	May 22 1987	545
11	3ct1	Ma Aye Thu	Loikaw	U Kyaw	June 19 1986	550
12	3ct2	Ma Hnin Yu	Kalay	U Min	May 23 1987	540
13	3ct3	Ma Sabai	Bamaw	U Aung	June 09 1986	548
14	3ct4	Ma Thazin	Hinthada	U Khine	December 04 1984	540

Table 4. Degree Table

No	Degree	CU	AY
1	B.C.Sc	Yangon	2007
2	B.C.Sc	Mandalay	2007
3	B.C.Tech	Yangon	2008
4	B.C.Tech	Dawei	2008
5	B.C.Sc(Q)	Yangon	1998
6	B.C.Sc(Q)	Yangon	1998
7	B.C.Sc(Q)	Taunggyi	2000
8	B.C.Sc(Q)	Yangon	2001
9	D.C.Sc	Bamaw	2001
10	D.C.Sc	Kalay	2003

5.3 Multidimensional Data Model

This section describes how to consider the data getting from data transformation process with multiple dimensions.

The results getting from attribute removing and record aggregation are considered with 2 data cubes each contain three dimensions, which is illustrated in Figure 2.

Time, University and Class dimensions are contained in data cube1 as illustrated in Figure A.

Time, University and Degree dimensions are contained in data cube2 as illustrated in Figure B.

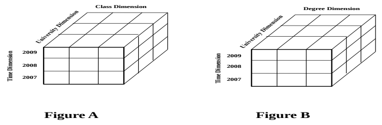


Figure 3. Multidimensional Data Model

5.4 Schema for Multidimensional Model

Data warehouse model can exist in the form of star schema, snowflake schema and fast constellation schema. According to the nature of domain area, star schema is chosen for schema representation. The following section is the detail explanation on schema representation.

5.4.1 Mapping between Schema and Data

In this section, how data are modeled onto star schema is explained.

(i) Star Schema for Data Cube1

In this schema, the student table is considered as fact table where fact value is number of student in Data Cube1. The other remaining tables become dimension tables. Therefore dimension table in Cube1 are Academicyear Dimension, University Dimension and Class Dimension.

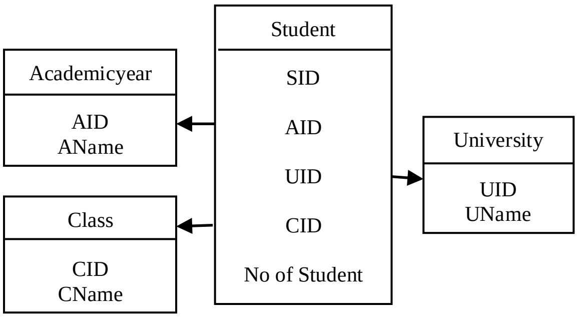


Figure 4. Star Schema for Data Cube1

(ii) Star Schema for Data Cube2

Similarly in Data Cube2, the fact table is student table and dimension tables are Academicyear, University and Degree tables.

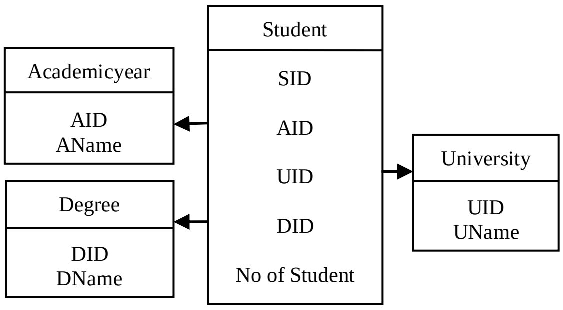


Figure 5. Star Schema for Data Cube2

6. Relational OLAP

This system uses Relational OLAP to store data into data warehouse. ROLAP architecture has two advantages: (a) it can be easily integrated into other existing relational information systems, and (b) relational data can be stored more efficiently than multidimensional data. ROLAP technology tends to have greater scalability than MOLAP technology [7].

ROLAP uses a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces. ROLAP includes optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services. It can handle large amounts of data.

6.1 OLAP Operation

This section describes the analysis output of our system implementation. This system can perform OLAP operations such as Slice/Dice, Drill down and Roll Up operations except Pivot operation.

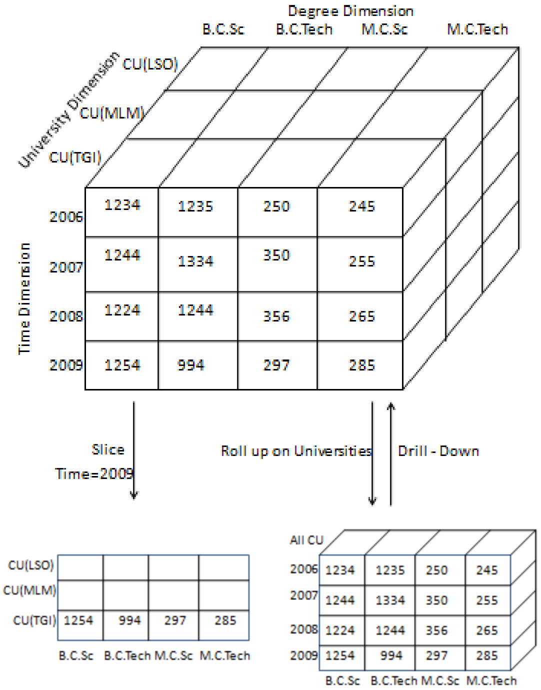


Figure 6. OLAP operation on Student Data Warehouse

(i) Slice/ Dice:

The slice operation performs a selection on one dimension (Time/University) on each Cube.

(For example, the number of student data (fact value) are selected for the Time Dimension using time= “2006”). Dice operation defines sub cube by performing a selection on two or more dimensions. (For example, the number of student data (fact value) are selected for the University= “CU(TGI)” and

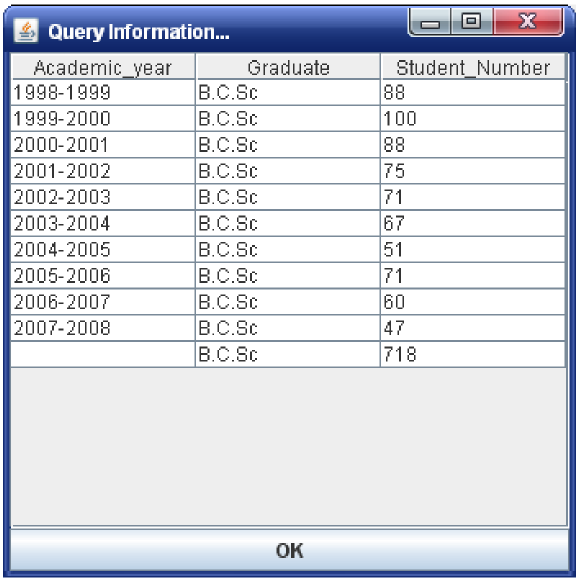
Time= “ 2009”). The sample for slice operation is illustrated in Figure 6.

(ii) Roll-up:

There are two concept hierarchies in our system. In time dimension, the concept hierarchy is semester <one academic year< all academic year. In University Dimension, the concept hierarchy is one University< all University in Myanmar. The roll-up operation performs aggregate function on these hierarchies. The sample for roll up operation is illustrated in Figure 6.

(iii)Drill-down:

Drill-down is the reverse of roll-up. Drill-down can be realized by either stepping down a concept hierarchy for a dimension. Figure 7 is the implementation output of drill down operation for Time dimension.



Academic_year	Graduate	Student_Number
1998-1999	B.C.Sc	88
1999-2000	B.C.Sc	100
2000-2001	B.C.Sc	88
2001-2002	B.C.Sc	75
2002-2003	B.C.Sc	71
2003-2004	B.C.Sc	67
2004-2005	B.C.Sc	51
2005-2006	B.C.Sc	71
2006-2007	B.C.Sc	60
2007-2008	B.C.Sc	47
	B.C.Sc	718

Figure 7. Result of drill down operation on time dimension

The system can query many information about information of student and university information depending on academicyear, university, degree and class with various dimensions. The system can show the total student number, pass percentage and student mark information with tables as well as bar chart according to the user chosen information.

7. Conclusion

Data warehousing is simply one component of modern reporting architectures. The real goal of reporting system is decision support to help people make better decisions. This paper focuses to enhance the processes on back end tools. In ETL, data set level transformation technique is used to transform data into student data warehouse. The student database is huge amount of data for all computer universities at all academic year. In this system, users analyze student information from student data warehouse using OLAP technique. This system can report information with table and bar chart. User can analyze number of student, pass percentage result and number of degree awarded student from student data warehouse. And the system can also add current student data and university information to be more effective and efficient into student data warehouse.

Reference

[1] Edgar, Oscar, Wolfgang Essmayr, Fraz Lichtenbeger, Werner Winiwarter, “An Authorization Model for Data Warehouse”, Center Hagenberg

[2] Fon Silvers,” Building and Maintaining Data Warehouse”

[3] Ham Kamber, “Data Mining Concepts and Techniques”

[4] M. Boehnlein, M .Plaha, A. Ulbrich, “Case Study-Buliding data for higher education in the course of microstrategy’s university program”

[5] M. Lupetin, “Data Warehouse Implementation using star schema”

[6] P. Vassiliadis,” A Survey of Logical Models for OLAP Databases”. ACM SIGMOD Record, December 1999

[7] Tara John, “Online Analytical Processing”, COCHI University of Science & Technology, August 2008