Key-Based Approach in Data Integration for Students Information System in Secondary School

Hasnah Nawang, Mokhairi Makhtar and Syadiah Nor Wan Shamsuddin

Faculty of Informatics and Computing, University Sultan Zainal Abidin, Malaysia

Abstract: With the emergence of ICT, the requirements for data integration have grown not only in business or banking, but in education as well. Students’ data contain important information that stores personal information, course details, examination information, scholarship status, co-curriculum data and many more. However the situation is getting complex when the data are stored in different systems and sources. A proper approach is needed in order to integrate those information based on user request. The purpose of this study was to integrate students’ data that exist from heterogeneous sources to ensure the accessibility by the user. The proposed method used was key based approach based on the primary and foreign key relationship. From the methods shown, this approach was proven suitable to be used in integrating and manipulating data for MRSM Kuala Berang, Malaysia.

Key words: Data integration • Heterogonous sources • Key based approach • Primary key • Foreign key and query

INTRODUCTION

Data integration is the process of combining data located in different locations and allowing the unified view called global schema [1, 2]. Data integration is also interpreted as the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information [3]. A successful data integration should able to deliver complete data from different sources to be stored in the data warehouse and must be easily accessed by the user.

There are many aspects to be considered in data integration such as designing, testing and techniques as well as the challenges in implementing it [4]. To solve the challenges in integrating data from different sources, a significant resource needs to be applied to understand not only the data and structures but also gathering and analysing the metadata of source systems to fully utilize the context of the information [5].

Data integration has become an active area in research due to the increasing sources from multiple system formats. Moreover, data integration also has been widely used in education. Educational data need to be accessible by students, teachers, school management and parents. Thus, this paper proposes a query based data integration in educational data for secondary school data warehouse.

Data integration becomes important in this research in order to access the sources from different formats and structures into a trusted unified view of school students’ data. In a real situation, students’ data are kept in different systems and different formats of heterogonous sources. Heterogeneity is defined as different and incompatible data and knowledge description formats exist due to legacy systems and the increasing variety of new approaches. Various types of heterogeneity can occur at many levels, including structural, syntax or semantic mismatch, access method, language, or protocol [6]. However, the challenge in data integration is to rearrange and match data from different systems in order to provide a unified information based on user request.

There are a few types of data integration approaches such as Manual Integration or Common User Interface, Application Based Integration, Middleware Data Integration, Uniform Data Access or Virtual Integration and Common Data Storage or Physical Data Integration.
Every approach is selected by each organization or individual based on the time consumed, how data will be represented to the users, the guarantee of the data quality or whether or not data integration needs a separate system to handle the fast volume of the data.

Some data sources do not have primary a key between them, which finally makes the identification of data changes between each file or source getting difficult. Furthermore, the data integrity also cannot be ensured when the data sources do not emphasise on the importance of primary key. Before the primary keys in each files are identified, the files or tables themselves need to be analysed in order to identify the dependencies of the tables. Then, this process will recognise the primary keys in order to create the relationships among sources.

In this research, the situation is to integrate data from students’ information system based on user queries that have been kept in 3 different sources. To integrate these files, these 3 schem as need to be matched to find the similarity between the data. In order to ensure the consistency of the data, the key based approach data integration are introduced.

**MATERIALS AND METHODS**

The implementation of data integration in this research involved sample data that were collected from Jun 2011 to March 2014 from form 4 and form 5 students of MRSM Kuala Berang. Students’ information system was used as a sampling data involving a total of 491 students’. Data contained in the students’ information system cover mainly on students’ personnel information, examination data and co-curriculum details.

**Data Extraction:** During the extraction process, the first step to be done is data cleaning for the purpose of maintaining the quality of the data. Data are analysed and incomplete data need to be removed before being integrated to reduce the amount of storage.

**Data Transformation:** The transformation step is required in generating surrogate keys. One of the minimum standards upon designing data, development and execution of data integration process is assuming all dimension tables have surrogate keys that function as the primary key [8]. Primary key is the unique identifier of an instance. On the other hand, a data integration key is defined as a set of data characteristics that could identify possible redundant physical data occurrences in a disparate data source [9].

During data transformation, data from different sources are analysed and converted into the same dimensions in order to support data integration. For this data integration, the Excel and csv format file need to be transform in the same dimension with the personnel and examination files stored in the database format. The step in transforming the files are explained below:

Step 1: Import the csv and Excel file to the database files.
Step 2: Change the type of the primary key or foreign key attribute in order to maintain the similarity of data types.
Step 3: Create a relationship between files using ‘KP’ and ‘NOMAK’.

The enforcement of the database structure such as the unambiguous primary key is one of the important processes at the transformation stage. For instance the primary and the foreign key cannot contain the null value, as an unhandled null value can destroy the integration process. Figure 1 below shows the primary key and foreign key of the tables from different sources and how the tables are joined using the keys.

![Fig. 1: The Primary and Foreign Key for Each Files](image)

From Figure 1, we can see the different primary keys in each table. In Personnel Table, the attribute set as the primary key is Acc No. Meanwhile for SPM01 table, the primary key is attribute named Nomak and for the SPMResult file, the primary key used is the attribute named AGiliran. However, the foreign key that exists in every table makes the integration easier by adding the relationship between the primary keys or primary keys and foreign keys. Foreign key is a field that is uniquely identified in a row of another table. For instance in this situation, KP and Nomak in Personnel have become the foreign key for Table SPM01 and SPMResult to ensure the integrity of the data.
Figure 2 shows how each table are connected using relationship and different primary keys in each table. Although each tables have their own primary key, the foreign key also functions as a connector to make relationship successful. In Personnel Table, the attribute that was set as the primary key is AccNo. For SPM01 table, the primary key is attribute named Nomak and for the SPMResult file, the primary key used is the attribute named AGiliran. However, the foreign key that exists in every table makes the integration easier by adding the relationship between the primary keys or primary keys and foreign keys. Foreign key is a field that is uniquely identified in a row of another table. For instance in this situation, KP and Nomak in Personnel have become the foreign key for Table SPM01 and SPMResult to ensure the integrity of the data.

**Query Implementation:** Queries were used in data integration to formulate the information request by user(Rayward et al. 2016). Query over a table is normal, but mostly queries require more than a table to display. For instance a school management wants to view the list of students who cannot graduate this year with the CGPA less than 2.00. This single query interface intends to view Name, Class and CGPA of the students less than 2.00 which involves rows from 3 different tables. This situation needs to retrieve the matching rows from 3 different tables. Besides, joined queries are usually used when user sets the limitation on what to display based on how the tables are connected to each other.

**RESULT AND DISCUSSIONS**

The significant main component in this research is the key element of the tables. Primary keys are a fundamental design feature of relational databases. Furthermore, a primary key is a combination of columns which are uniquely specified in a row [10]. Every table contains its own primary key, however the existence of the Foreign Key helps in integrating the files to avoid the redundancy of data. For example in the above query,
AccNo is the primary key for the Personnel table, however, AccNo does not exist in the Graduation and SPM04 tables. Seems the Foreign Key Nomak exists in the Personnel tables and also in Graduation and SPM04 as the primary key, the query can match the table using that key.

During the research, mostly, selective operation queries were used by joining the tables using primary key and foreign key. Inner join was expected to produce query result by combining the rows between 2 tables to match up the tables based on the specific criteria given by the user. In Section 5’s query example, the criteria is CGPA < 2.00 and Grad_Status = Fail. The row or attribute that functioned as a match is the Nomak in the 3 tables. This means that if there exists a row from the first table in that matches rows in the second table, then the query will be returned in the results based on user requirements.

Besides, query also plays an important role in this research as it will perform different functions depending on the user request. The main important role of queries is to retrieve specific data from a table or more. After the query has been successfully executed, the queries need to be mapped to the global schema. Figure 5 shows the proposed framework of the integration and mapping process in this research.

In order to evaluate the results of the implement approach, we will look at the data shown in Table 1 based on query example discussed in the Query Implementation section. From the records, we can see the list of the executed query from 3 different files as shown in Figure 4, which has been transformed in the same dimensions. Matching parameters that have been used as a relationship has become the mechanism in order to ensure the query results are not duplicated.

### Table 1: Global Schema Table

<table>
<thead>
<tr>
<th>AccNo</th>
<th>Name</th>
<th>Class</th>
<th>Grad_Status</th>
<th>CGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>11002</td>
<td>Aminah</td>
<td>5E</td>
<td>Fail</td>
<td>1.89</td>
</tr>
<tr>
<td>11004</td>
<td>Sarah</td>
<td>5F</td>
<td>Fail</td>
<td>1.53</td>
</tr>
</tbody>
</table>

### CONCLUSIONS

Data integration has become important in the educational field. In this paper, we aim to do the data integration for the educational data to be easily accessed by the user. The sample has been collected from MRSM Kuala Berang since 2011 to 2014. The proposed technique is Key-Based integration using Primary and Foreign Key as the key point. We have explored the features and functions and provided the example of SQL queries that show how data can be integrated.

The finding is useful for current and future work in integrating and manipulating students’ data in MRSM Kuala Berang, Malaysia. However, this paper should be further enhanced in future with a big set of data that can be collected from other schools in order to get the test the effectiveness of the proposed method in a large amount of data.

### REFERENCES