Transformation Rules for Model Migration in Relational Database Preservation

Arif Ur Rahman
Faculdade de Engenharia,
Universidade do Porto
INESC Porto
badwanpk@fe.up.pt

Gabriel David
DEI - Faculdade de Engenharia, Universidade do Porto
INESC Porto
gtd@fe.up.pt

Cristina Ribeiro
DEI - Faculdade de Engenharia, Universidade do Porto
INESC Porto
mcr@fe.up.pt

ABSTRACT

Digital preservation is about memory and giving easy access to it. If the digital object is a relational database the requirements of normalization may make it hard to access and understand. In order to deal with this problem we have proposed the DBPreserve approach to transform a relational database to a dimensional model as part of the preservation process, making the preserved information more explicit and easier to access. The paper presents a set of transformation rules to deal with aspects of the migration process such as the identification of the fact tables corresponding to the main organizational processes and the choice of the set of relevant dimensions. The rules help to keep the traceability of the migration process and to preserve integrity and authenticity. The rules were implemented in a case study which involved a human resources information system.

Keywords

Database preservation, database transformation rules

1. MODEL MIGRATION

A relational database incorporated in real information systems normally has a complex structure, integrity constraints, triggers, functions stored procedures and applications developed in high-level language. This makes preserving and using the information in the future difficult. The DBPreserve approach proposes a solution for preserving relational database systems for the future [1]. A relational database is migrated to a dimensional model to make it simple to understand as well as easy to access. In the process of migration, data transformations are needed as changes may occur in the structure and representation of data.

Dimensional modeling is a logical design technique that seeks to present the data in a standard framework which is intuitive, allows for high-performance access and is resilient to change. The strengths of dimensional modeling make it a better choice for long term preservation and access of information.

In the process of migration the information embedded in code is calculated and explicitly stored. This makes the dimensional model independent of the DBMS details and application logic. In the sequel we propose a set of transformation rules which help to effectively carry-out the migration process.

2. TRANSFORMATION RULES

Implementing the transformation rules corresponds to concrete extraction, transformation and loading processes where data is selected, cleaned, formatted accordingly, checked for referential integrity against dimensions and transferred. The rules can be grouped into categories. They include:

1. **Generic Table Information**
   (a) Description of each table is prepared including table level information such as name of the table, number of rows, number of columns and a short description. Furthermore, the description also contains column level information such as column names, number of nulls, data type, a short description, distinct values, minimum and maximum values in the column.
   (b) Tables with no data are analyzed and if there is no need to keep them, they may be ignored. In each table there may be some columns which are empty for all the rows. They may also be ignored.
   (c) Sometimes snapshots of tables are taken on a specific date and kept in the database. Also, there may be tables which store data as a preliminary step for all or part of it to be included in a database. If tables are found to be of this nature, they may also be ignored.
2. **Keys**
   (a) The primary keys of tables need to be recorded. Primary keys should be known for the tables which are potential candidates for dimensions though it is not necessary for all tables.
   (b) Foreign keys in tables are also recorded. Situations may arise where there maybe orphan child

iPRES2011, Nov. 1–4, 2011, Singapore.

records in tables. Orphan child records need special attention in the migration process in order not to lose them.

3. **Processes, Facts and Dimensions**

   (a) Tables are clustered considering their foreign keys and under the broader context of the relevant processes in the organization. The organizational processes about which the system stores data are found out.

   For each process the set of participating tables is listed. Furthermore, in each set of tables there is normally a central table which is identified. Usually, it has no incoming references but it references other tables. The central tables have the real world facts recorded in the organizational processes. They may be candidates to be loaded into fact tables and become the centers of stars.

   (b) Dimensions involved in each organizational process are identified. Analyzing each cluster of tables and having identified which tables are likely to be the sources for the fact tables in the future stars, the remaining tables are candidates to be the source of dimensions. However, their identification is more accurate if it is guided by the knowledge of the organizational processes and of their main entities. Once all the dimensions are identified a bus-matrix is constructed.

   (c) The migrated model should be made of simple stars, to be easy to query. One technique to achieve this is to de-normalize the dimensions, including simple or multiple hierarchies in each one of them. This corresponds to merging tables in the original model.

   (d) There may be situations where a set of tables in the operational system may need to be joined for constructing a dimension and one of them is a lookup table with more records than actually used by the lower level in the hierarchy. In such situations a snowflake schema is constructed to keep the higher level rows.

4. **Nulls**

   (a) In the process of migration if nulls need to be replaced, they should be replaced with a value which has no meaning in the domain.

5. **Code**

   (a) In a database system the application program typically has forms for adding new data, displaying the data already in the system and generating reports. Screen shots of the forms are taken and preserved.

   (b) **Short description of algorithms (code)**

   If there are functions, procedures or code in any form to derive information from the data stored in a database, they are executed and the results are explicitly stored. Furthermore, a description of each piece of code explaining the code and the information it produces is written and kept in the preserved database.

The mappings between the original and the migrated models, which is the base of the ETL process, must be kept as preservation metadata to document the whole process, remain as evidence of the data origin, and facilitate any verification procedure.

3. CASE STUDY

The transformation rules presented in Section 2 were used in a case study. It involved the human resources information system of a higher education institution. The database stores all the information required by the institution to manage the information on teachers and the administrative staff.

In the case study a mapping between the original and the migrated models, which are the base of the ETL process was developed. This mapping was kept as preservation metadata to document the whole process, remain as evidence of the data origin, and facilitate any verification procedure. The information gathered according to rule number 1 helps in performing a completeness check on the data. The recording of keys in rule number 2 makes clustering tables easy. For identifying the organizational processes about which the system stores data, an analysis of the application software used to interact with the database was very helpful. `Contract` is the main organizational process about which the system stores data. Each time a new employee is hired, promoted, assigned extra duties or retired, the data is recorded by this process. In the dimensional model the fact table stores information like the hire date, the contract renewal or expiry date, monthly salary, duration and so on. The fact table is surrounded by dimensions which store the biographical data, the cadre, the unit where an employee works and a date dimension.

In the migration process null values needed to be replaced in columns of type date and character. In the date column the nulls were replaced by 01 Jan 0001 and 31 Dec 9999 depending on the situation and in character type columns the nulls were replaced by 'Unknown'.

The users of the system were involved in preserving the database which helped in carrying-out the task. The migration process also gave the opportunity to detect any missing information and wrong information. It was notified to the users and was corrected.

Although sharing with traditional data warehouse systems many intuitions and techniques, the ultimate goal of preserving a database is very different from the usual goal of building a decision support system. This has some consequences in the nature of the fact tables, which often lack clear measures or the measures included are just secondary elements.

4. REFERENCES

[1] A. U. Rahman, G. David, and C. Ribeiro. Model migration approach for database preservation. In *International Conference on Asian Digital Libraries (ICADL)*, volume 6102, pages 81–90, Springer-Verlag Berlin Heidelberg, 2010.