Using Dimensional Modeling Techniques in Data Warehousing

Alexandra Maria Constantin[1,2], Ionut Teodor Ionescu[1], Serban Stancu[2], Nicoleta Ruse-Constantin[2], Florentin Catalin Neacsu[2], Andrei Daniel Ionescu[2], and Andreas Mihai Toma[2]

[1]The Bucharest University of Economic Studies, ROMANIA

[2]BI Company, ROMANIA

Abstract: *Starting from the fundamental concepts of Data Warehousing and their importance in Business Intelligence applications, we may speak of the necessity of multideminsionalizing data and their relations, in the sense that the transaction processing system, through its complexity, attracts an exponential development of techniques involved in extracting information from data, a supporting process in decision making. In the virtue of rapid development of data warehousing techniques in extracting information from data and abstracting existing relations between data structures, the current paper aims to: present an introduction in approaching data warehousing, and to present two specific approaches in current data warehousing and presenting the conclusions.*

Keywords: *data warehouse, modeling techniques, surrogate key, ETL, BI Solutions*

1. Introduction

Starting from the fundamental concepts of Data Warehousing and their importance in Business Intelligence applications, we may speak of the necessity of multidimensionalizing data and their relations, in the sense that the processing system of transactions, through its complexity, attracts an exponential development of techniques involved in extracting information from data, a supporting process in decision making.

Classic methods of data querying, taken through the process of normalization, do not ensure a high degree of accuracy in extracting information from the dataset and in optimizing existing connections between data, being encountered in situations in which all error cases cannot be efficiently accounted for in the operational database. As information, owned by organizations in large volumes, can exist along multiple platforms, in different storage environments, (operationally developed database) and under different structures, the necessity of Data Warehousing processes becomes apparent.

The process of data warehousing is focused [1] on: implementing the afferent processes of optimally accessing data in the operational database; storing the data in a structure that allows for high access and data transformation flexibility, obtained through advanced filtering of high volume operational data.

Moving on to the concept of Data modelling [1], it designates the process of abstracting data, thus allowing for a concise interpretation of business reality. To continue, we will focus on presenting some of the more important aspects of data modelling techniques, such as data restructuring based on surrogate keys and natural keys.

2. Dimensional model techniques

Dimensional models are not affected by changes in the data connections, caused by modifications to the data set. In other words, interrogations on a BI level and built applications will not be impacted when new modifications are implemented in BI solutions. To ensure feasibility in the data to be used to take decisions, the necessity of implementing the surrogate key in tables appeared, starting from the premise of notable probability of corrupt data existing in the operational database. Thus, the surrogate key, used as a primary key, is automatically generated in the tables already brought to a minimal form for implementation in ETLs. Also, the existence of a surrogate key brings the following benefit: one can track the history of modifications to the data.

The problem of dimensional modelling does not stop at the level of primary keys, it also continues with difficulties encountered [2] in the design and maintenance phase, which reflects in BI Solutions structures and in existing connections in the reporting level.

http://dx.doi.org/10.17758/ER715211

Starting from the example of document flux in an organization, in which financial documents are subordinated to contracts and subcontracts, which in turn can be subject to other contracts, we may say that we are dealing with a hierarchic structure. The problem posed is identifying the document in a chain of documents, subordinated to contracts, by including the hierarchic structure in the dimensional model. The structure of a document is presented as follows:

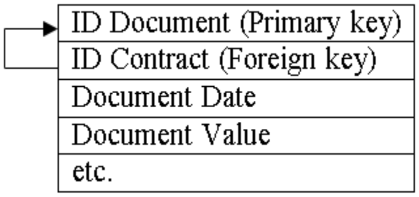

Fig. 1: The basic structure of a document in the document flow.

Starting from the structure presented beforehand, we may notice an important aspect: the contract is, itself, a document. From this, we may deduce that we can situations of self-referencing. A decent solution in manipulating the parent-child hierarchy is given by the OnLine Analytical Processing solution, in short, OLAP, which offers multiple advanced interrogation tools. However, there are certain difficulties where DML (Data Manipulation Language) is concerned, in the sense that not all SQL platforms support advanced DMLs.

A solution for the proposed hierarchic structure is given by using surrogate keys in designing tables in an ETL process (Extract, Transform and Load). The new approach for surrogate keys is:

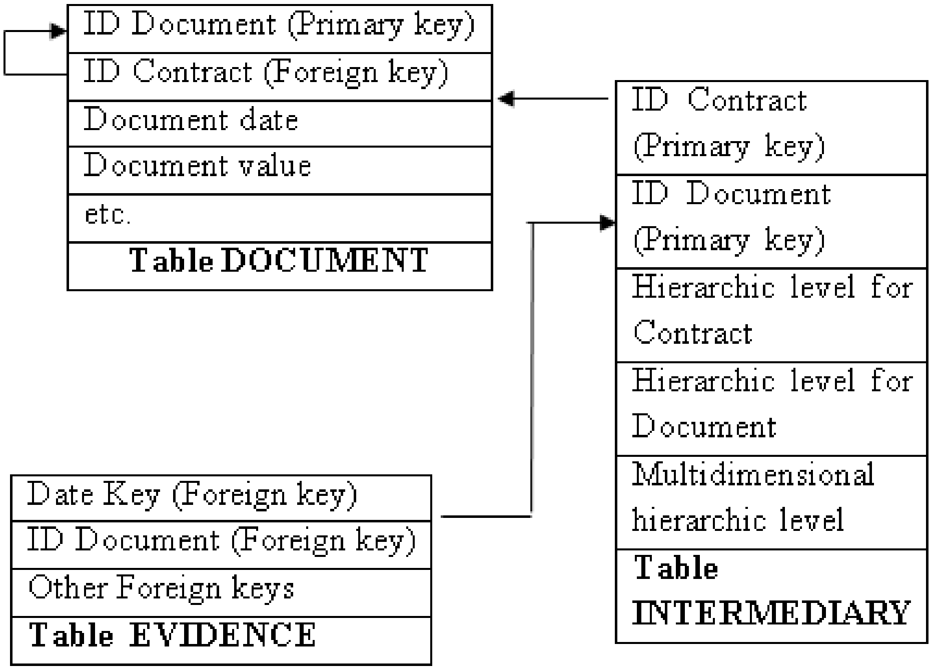
ID Document (Primary key)
ID Contract (Foreign key)
Document date
Document value
etc.
Table DOCUMENT
ID Contract (Primary key)
ID Document (Primary key)
Hierarchic level for Contract
Hierarchic level for Document
Multidimensional hierarchic level
Table INTERMEDIARY
Date Key (Foreign key)
ID Document (Foreign key)
Other Foreign keys
Table EVIDENCE

Fig. 2: The relationship structure of documents.

http://dx.doi.org/10.17758/ER715211

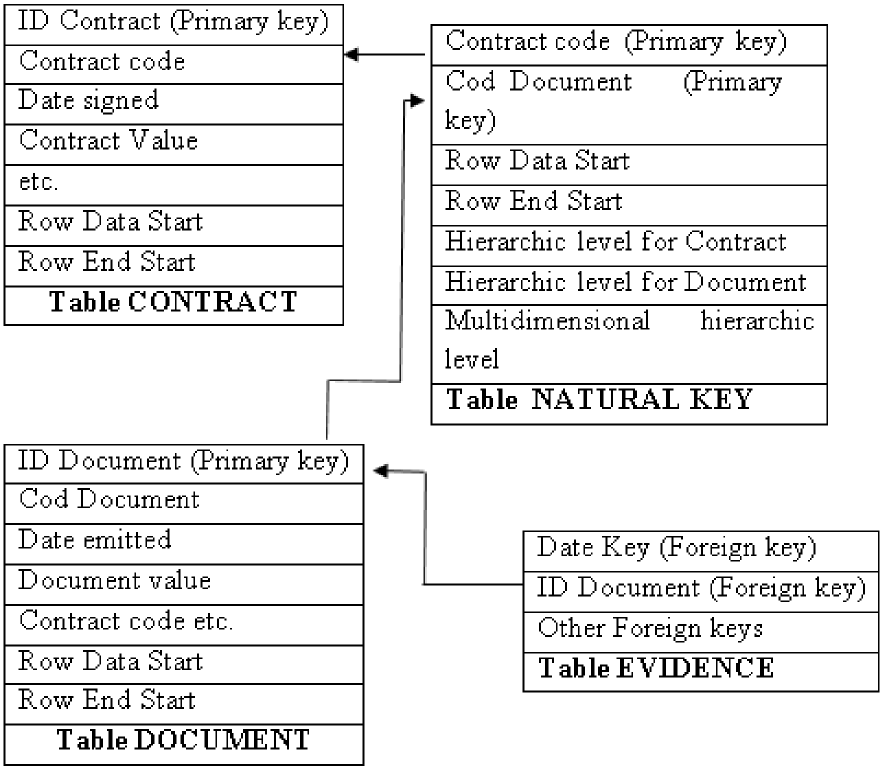

Fig. 3: The new form of relationship structure of documents, derived from structure presented in figure 2

The relationship structure presented in figure 2 helps diminish identification time of information about ongoing documents, in a certain order, or about those that have passed in archiving, at the time of completion of connected projects. Also, it is worth mentioning that the table "Intermediary" is meant to generate hierarchic connections between documents.

The hierarchic level for contract corresponds to situations when documents are directly linked to the "parent" Contract. The hierarchic level for Documents is given by situations when the connection is "document-to-document", which is to say a document is connected to another document which is in turn connected to another document or a subcontract. The multidimensional hierarchic level corresponds to a combination of previous levels, in the sense that, at a given time in the document flow, the necessity of mapping a document situated in the Contract hierarchic level with a document situated in the Document hierarchic level develops.

However, the "intermediary" table presents some drawbacks, such as [2]: difficulty in construction, performance problems in interrogation due to large number of rows, a high degree of limitations whereas using it ad-hoc is concerned, etc.

The necessity of optimizing maintenance and following the history of modifications in the representative data (not errors) has led to modifications to the structure described by figure 2, by eliminating the surrogate key. In the relationship structure of document hierarchy from figure 2, a new table was introduced, named "Natural Key", which brings the following advantages [3]: reducing the number of dimensions to one, both in the contract and document levels; inclusion of two key elements: start date and end date for every row (in order to improve logging of modifications suffered by the data set) etc.

3. Conclusions

Efforts in improving data warehousing processes must be focused, as much as possible, on using advanced multidimensionalizing techniques, so as to ensure an as-high degree as possible in complex queries.

On the other hand we must take into account that current approaches in data warehousing, two of which were presented in this paper, ensure, in an abstract way, a rendition of reality, but not to the same extent, data consistency.

4. References

[1] Ballard, C., Herreman, D., Schau, D., et al. (1998*) Data Modeling Techniques for Data Warehousing, International Technical Support Organization* , IBM Corporation, California

[2] Casters , M., Bouman , R.,, van Dongen, J. (2010) *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*, ISBN: 978-0-470-63517-9, Wiley Publishing

[3] Imhoff, C., Galemmo, N., Geiger, J. (2003) *Mastering Data Warehouse Design: Relational and Dimensional Techniques*, ISBN: 0-471-32421-3, Wiley Publishing

[4] Kimball, R., Ross, M. (2013) *The Data Warehouse Toolkit – Third Edition*, pp. 146-148, John Wiley & Sons Inc, ISBN: 978-1-118-53080-1, Indianapolis

[5] Mundy, J. (2009) *Five Alternatives for Better Employee Dimension Modeling*, *Kimball Group*, available at: http://www.kimballgroup.com/2009/08/five-alternatives-for-better-employee-dimension-modeling/

[6] Provost, F., Fawcett, T. (2013) *Data Science for Business: What you need to know about data mining and data-analytic* thinking, ISBN:978-1-449-36132-7, O'Reilly Media

[7] Westerman, P. (2000) *Data Warehousing: Using the Wal-Mart Model (The Morgan Kaufmann Series in Data Management Systems),* ISBN: 1-55860-684-X, Academic Press

http://dx.doi.org/10.17758/ER715211