Goal-Driven Design of a Data Warehouse-Based Business Process Analysis System

LAILA NIEDRITE, DARJA SOLODOVNIKOVA,
MARIS TREIMANIS, AIVARS NIEDRITIS
Department of Computer Science,
University of Latvia,
Raina bulv. 19, Riga, Latvia.

Abstract: In this paper we propose an approach to business process measurement and control. The process measurement is accomplished by a data warehouse. We propose a method for development of a conceptual model of a data warehouse based on GQ(I)M method to define business goals. Together with definition of business and measurement goals, entities and their attributes are identified. The model composed of these entities is designed with UML 2.0 structure diagram in accordance with our proposed indicator determination metamodel. The indicators are identified in compliance with GQ(I)M method and defined with OCL expressions using entities and attributes of the model. Then, based on the structure of the OCL expressions, potential facts and dimensions of the data warehouse are identified. The implementation of the proposed solution is discussed for the university data warehouse project.

Key-Words: Data warehousing, Modeling, Goal-driven methodology

1 Introduction

Every institution attempts to organize its activities in the most effective way. To improve business processes, they first should be understood. One of the well-known definitions is the following [1]: "process is any activity or group of activities that takes an input, adds value to it and provides an output to an internal or external customer. Processes use an organization's resources to provide definitive results". So business processes ensure achievement of the institution's goals. To make it possible to determine whether goals are achieved, it is necessary to measure these processes, and the measurements should be compared with the target values to make some conclusion and if required to change the business processes. In [1] the author declares: "Measurements are the key. If you cannot measure it, you cannot control it. If you cannot control it, you cannot manage it. If you cannot manage it, you cannot improve it." The definition of the measurement process is the following [2]: "Measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to characterize the attributes by clearly defined rules."

Corresponding to the time and purpose of the measurement, process monitoring and process measurement [3] can be distinguished. *Process monitoring* means supervision of the situation to make a conclusion. Data from a process execution system, for instance, workflow system log files, are used in real-time. Process data are collected according to indicators. *Process measurement* means recognition of the business process execution results using indicators. Process measurement combines data about processes with business data. In [3] several implementation possibilities for process measurement and monitoring are distinguished: log file analyzers, data warehouse, which can be used in case when business data are already available in the data warehouse, as well as the process measurement and monitoring tools.

Data warehouse as a solution for business data storage is discussed in many papers. In [4] the Process Data Warehouse is defined „as a data warehouse which stores histories of engineering processes and products for experience reuse, and provides situated process support". The concept of Performance Management System is defined in [5] as the system, which "stores and manages all performance relevant data centrally, including both financial and non-financial data", and also ensures systems approach to measurement and timely access to data. In [6] the authors propose Corporate Performance Measurement System, where process performance data are integrated with the institution's data warehouse. As the data source for internal business process characterization, the log file of the workflow system is used.

Section 2 of this paper is dedicated to the case study business process examples. In Section 3 the method to obtain indicators for the process measurement is presented. Section 4 introduces the method to derive a data warehouse model from indicator definitions with Object Constraint Language (OCL). In Section 5 we discuss the implementation of the process monitoring and measurement system. In Section 6 we present related work. We conclude with directions for future work in Section 7.

2 Case Study

Further in the paper we will discuss two closely related workflows at the university as an example that illustrates the application of our method. *Students' enrolment in courses* is a workflow, when students apply for courses through internet or face to face with study advisors. Students with financial or academic debts can not apply for courses through internet. Students can also cancel their applications for courses. Students must confirm their course selection to be enrolled in a course, but it is possible to cancel enrolment until the end of the enrolment period. The process participants were not satisfied by the organization of the process, therefore, it was decided to carry out the process measurement to identify the necessary changes. *E-learning* is a workflow, when students are automatically registered for an e-course, if students are enrolled in the course and this course exists in the e-learning environment. During the semester students use e-course materials, however e-course instructors support students' learning process. The development of e-courses was financed within the framework of the e-university project, therefore it was important to examine the issues related to the usage of developed courses.

3 The Method and Metamodel for Indicator Determination

The determination of indicators necessary for the process measurement was based on the goal-question-indicator-measure GQ(I)M proposed by [7], which extends Basili's work about GQM [8]. GQ(I)M method of [7] differs from the classical GQM in indicator determination.

The following elements of GQ(I)M method are necessary to identify indicators – Mental Model, business and measurement goals and indicators. We renamed the Mental Model to the *Notional Model* in the metamodel (Figure 1) that corresponds to our approach, because in our opinion this name better conforms to the contents of the model. In order to describe the mutual relations between indicators and attributes, in the indicator determination metamodel we have introduced the additional association between the classes *Indicator* and *Attribute* and the association class that depicts the *Transformation Function*. In our method we do not use the measure concept, because we assume that constructing the Notional Model, it is possible to define all interesting properties of each measured entity at the attribute level, including measures according to GQ(I)M interpretation.

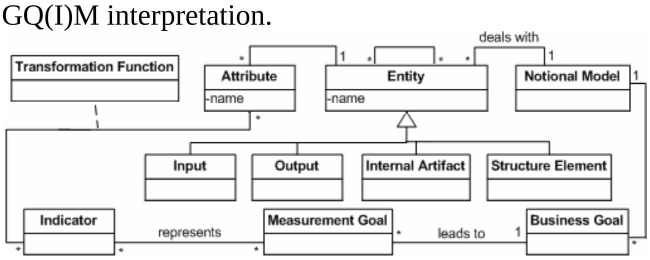

Fig. 1. Indicator determination metamodel with UML 2.0 [9]

Initially the strategic *Business Goals* are identified. The Notional Model that incorporates *Entities*, about which some information is necessary, is designed. There are 4 types of Entities – *Input* (resources), *Output* (products, effects), *Internal Artifacts* (data, tools, knowledge) and *Structure Element* (subprocesses, activities, flowpaths). Each Entity is characterized by *Attributes*. Entities can be interconnected. This model is designed during the goal-driven measurement process interviewing stakeholders of a business process to identify Business Goals. Simultaneously, the information about Entities and their Attributes managed and affected by stakeholders is gathered. Then *Measurement Goals* are determined. The questions are put to identify quantitative measurement results together with their presentation logic, which are called *Indicators*. Basically an Indicator is information about a process that helps to ascertain the achievement of the goal. The indicator definition contains data elements that are not depicted in the metamodel, because its purpose is to explain the mutual relations of Indicators and Attributes at the Indicator level.

3.1 Indicator Determination for Case Study

According to the previously explained method, we identified two Business Goals for the case study correlated processes: (1) improve the students' enrolment process and (2) provide efficient and effective e-learning.

Table 1. Questions and Indicators

Goal G1			
Question		Indicator	
2	What was the shortage of courses?	I3	Cases when applicants did not have the opportunity to enrol into courses
		I5	Number of language courses with big number of applicants
		I6	Number of students enrolled in a course among all applicants
3	What is the cancellation procedure of courses with small number of applicants?	I8	Number of cancelled courses
4	How many students could not enrol in courses through internet and why?	I10	Number of students with financial debt
		I11	Number of students with academic debt
5	Was the course offering duly prepared?	I12	Course offering at the beginning of enrolment
Goal G2			
Question		Indicator	
6	How many students enrolled in courses through internet?	I17	Number of part-time students, who applied for each course through internet
8	How did the workload of study advisors change?	I22	Number of students enrolled in a course by study advisors
10	Analysis of workflow activities by days	I26	Number of cancellations for each course
		I27	Number of enrolment activities in each course
		I28	Number of cancellations of enrolment in each course
Goal G3			
Question		Indicator	
11	What is the result of investment according to accessibility?	I30	Investments per department
		I31	Enrolment in e-courses per department per time
		I32	Number of e-course students per department per time
12	What is the result of investment according to usage?	I34	Number of hits of students per department per time
		I35	Usage time of students per department per time
13	What is the result of investment according to student?	I36	Number of e-course students' sessions per department per time

Then Business Goals were analyzed and detailed until subgoals. As a result a number of Measurement Goals were identified. We will further discuss three of them:

G1. Improve the effectiveness of enrolment process from the students' viewpoint.

G2. Improve the efficiency of enrolment process and related processes from the viewpoint of the academic department.

G3. Improve the efficiency of the investment in e-learning from the viewpoint of the top management.

After the definition of Measurement Goals, questions that characterize achievement of the goals were formulated and indicators that answer these questions were identified. Altogether 13 questions and 36 related indicators were determined. Because of space limitations, in this paper we will include only such Indicators (Table 1) that are essential to demonstrate the derivation of the data warehouse model.

3.2 Design of the Notional Model

Together with definition of goals, entities and their attributes that are essential for the operation of the organization are identified.

The notional model composed of these entities and attributes is designed with UML 2.0 structure diagram in accordance with our proposed indicator determination metamodel. The extended UML 2.0 structure diagram was also presented in [10] to model the business context of the process. The authors are using OMG [9] class definition, where class is "a set of objects that share the same specifications of features, constraints, and semantics. The purpose of a class is to specify a classification of objects and to specify the features that characterize the structure and behaviour of those objects." Therefore, classes can be used to model business processes.

Corresponding to GQ(I)M method, also the process context is interesting in the Notional Model, but in a different aspect, in accordance with the indicator determination metamodel, therefore we are inspired by [10] solution.

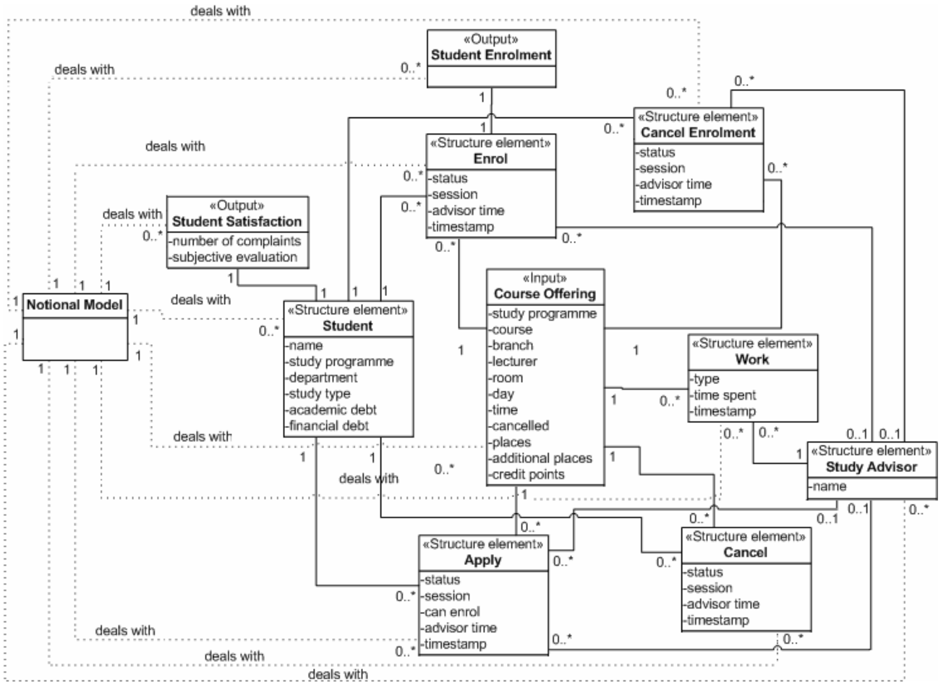

Fig. 2. Notional Model of students' enrolment in courses

We developed the Notional Model for each measured process used in our case study. Figure 2 illustrates students' enrolment process, but Figure 3 demonstrates the Notional Model of the e-learning process. To improve understandability we used dotted line to represent associations between Notional Model and Entities in Figures 2 and 3.

Since these processes are correlated, both models include several common elements – Entities *Student* and *Student Enrolment*, which change their type in the second process, for example, *Student Enrolment* changes from Output to Input.

3.3 Formulating Transformation Functions with OCL

When the Notional Model is designed, Transformation Functions of Indicators identified according to GQ(I)M method are formulated with OCL query operations [11] that return a value or set of values using Entities, Attributes and associations from the Notional Model. The Indicators from Table 1 are formulated with OCL in Table 2.

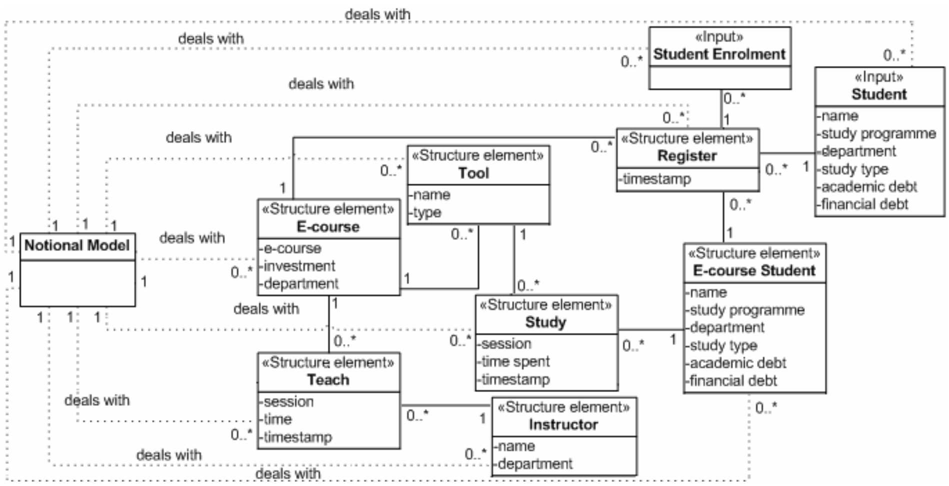

Fig. 3. Notional Model for the E-learning Process

Table 2. Indicator formulation with OCL

I3	**context** Notional Model::I3():Integer **body:** Apply->select(status='Applied' and can enrol='No')->size()
I5	**context** Notional Model::I5():Integer **body:** Course Offering->select(Apply->select(status='Applied')->Student->asSet() ->size()>places and branch='Language courses')->asSet()->size()
I6	**context** Course Offering::I6():Real **body:** (Enroll->select(status='Enrolled')->Student->asSet()->size())/(Apply-> select(status='Applied')-> Student->asSet()->size())
I8	**context** Notional Model::I8():Integer **body:** Course Offering->select(cancelled='Yes')->size()
I10	**context** Notional Model::I10():Integer **body:** Student->select(financial debt='Yes')->size()
I11	**context** Notional Model::I11():Integer **body:** Student->select(academic debt='Yes')->size()
I12	**context** Notional Model::I12():Set(Course Offering) **body:** Course Offering
I17	**context** Course Offering::I17():Integer **body:** Apply->select(Study Advisor->isEmpty())->Student->asSet()->select(study type='Part-time')->size()
I22	**context** Course Offering::I22():Integer **body:** Enrol->select(status='Enrolled' and Study Advisor->notEmpty()).Student->asSet()->size()
I26	**context** Course Offering::I26():Integer **body:** Cancel->size()
I27	**context** Course Offering::I27():Integer **body:** Enrol->size()
I28	**context** Course Offering::I28():Integer **body:** Cancel Enrolment->size()
I30	**context** Notional Model::I30(f:String):Real **body:** E-course->select(department=f).investment->sum()
I31	**context** Notional Model::I31(f:String, t:Date):Integer **body:** E-course->select(department=f).Register->select(timestamp=t)->size()
I32	**context** Notional Model::I32(f:String, t:Date):Integer **body:** E-course->select(department=f).Register->select(timestamp=t)->E-course Student->asSet()->size()
I34	**context** Notional Model::I34(f:String, t:Date):Integer **body:** E-course->select(department=f).Study->select(timestamp=t)->size()
I35	**context** Notional Model::I35(f:String, t:Date):Integer **body:** E-course->select(department=f).Study->select(timestamp=t).time spent->sum()
I36	**context** Notional Model::I36(f:String, t:Date):Integer **body:**Study->select(Tool.E-course.department=f and timestamp=t).session->asSet()->size()

4 Deriving Dimensions and Facts

OCL query operations that define Indicators are further analyzed to design a data warehouse model. Firstly potential facts are identified. If a result of an operation is numerical, for example, sum(), size(), round(), multiplication, division, we consider such values as potential facts. From the data in Table 2 we have identified the following potential facts of the case study data warehouse: Apply->size(), Student->size(), Course Offering->size(), Cancel->size(), Enrol->size(), Cancel Enrolment->size(), E-course.investment->sum(), Study.time spent->sum(), Register->size(), E-course Student->size(), Study->size(), Study.session->size().

Secondly potential dimensions and dimension attributes are determined. Initially classes, which appear in *context* clause of OCL query operations excluding the class Notional Model, are considered as potential dimensions. Their attributes correspond to dimension attributes. For instance, the dimension Course Offering (Figure 4) was obtained in that way from the OCL operations showed in Table 2. In addition other dimension attributes are derived from class attributes used in *select* clause of OCL query operations. These attributes are grouped into dimensions corresponding to classes that contain these attributes. For example, the following additional dimension attributes were obtained from the operations given in Table 2: *status*, *can enrol* of the class *Apply*; *study type, financial debt, academic debt* of the class *Student*; *department* of the class *E-course*; *timestamp* of the class *Register*; *timestamp* of the class *Enrol*. When dimensions are identified, some class attributes from the Notional Model may be missing in a data warehouse model. These attributes can show the new directions for analysis that have not been discussed before. These attributes can be added to the data warehouse model after interviews with decision makers.

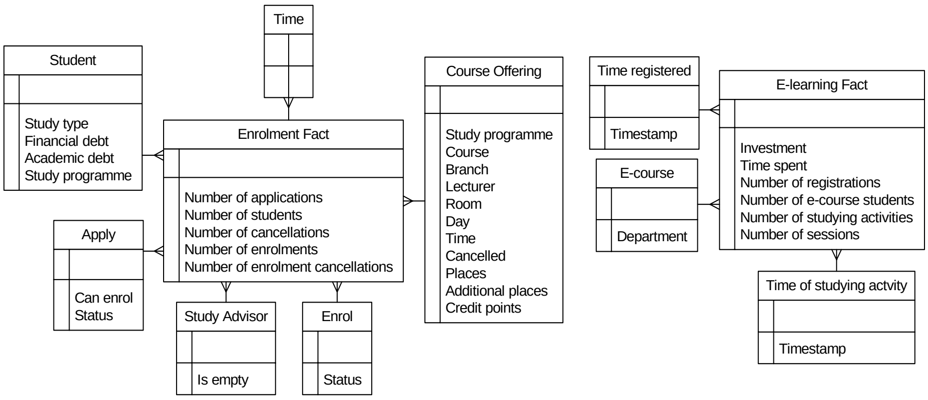

Fig. 4. Data Warehouse Model

Finally the time dimension is added to the data warehouse model.

Two data warehouse models (Figure 4) were produced for the case study indicators. The potential fact Course Offering->size() was not included in the model, because it will always equal to 1.

5 Implementation of the Process Monitoring and Measurement

After development of the data warehouse model and its mapping with data sources we discovered that some data are not available. In order to accumulate the data about Structure Elements from the Notional Model that correspond to process activities, such as *Apply*, *Enrol*, records in a log file were created for each workflow execution.

In our case study during the implementation of the students' enrolment process, the procedures of the operational information system were implemented in such a way that each call of the related procedure was fixed in the column PROC of the special log file table USERLOG with the format <procedure_name>(<course_code>). This table also consists of other columns: date, user login, number of activities in the day.

The workflow of the e-learning process is reflected in the e-learning management system's log files. A new web server log file is created each day and is loaded into the USERLOG table. The information about activities is obtained analyzing URL.

The whole process measurement and monitoring scheme is illustrated in Figure 5.

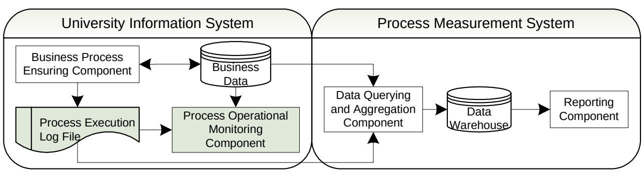

Fig. 5. Process measurement and monitoring

The process monitoring elements are depicted in grey. The *Process Operational Monitoring Component* creates separate reports from the log file data and from the business data. The log file data and business data are integrated and analyzed by the *Process Measurement System* based on the data warehouse models, which were designed according to the earlier presented method.

The Process Operational Monitoring Component supports analysis of Indicators about process workflow directly from the log file during the process execution. Then Indicators are loaded into the data warehouse and used for the process measurement.

6 Related Work

Similar approaches, when goal-driven requirements analysis is proposed together with the method to obtain the conceptual model of a data warehouse, which stores indicators required for the goal achievement support, are presented in [12] and [13].

In [12] the conceptual model of a data warehouse is designed using a logical diagram obtained during decisional modelling. Facts are mapped onto entities or relations of data sources. Hierarchies of each fact are constructed by path of many-to-one associations in the data source in the same way as in [14].

In [13] the method is composed of three steps: top-down analysis, bottom-up analysis and integration. The authors also use GQM approach for top-down analysis to complete abstraction sheets. Data from abstraction sheets are used to design an ideal star schema for each goal. During the bottom-up analysis candidate star schemas that really can be implemented from available data sources are identified. The last step is integration, when ideal requirements are matched with the real schema.

In [6] the authors propose the Corporate Performance Measurement System, where process performance data are integrated into the enterprise data warehouse. The authors present a method to design a data warehouse model, when specific goals of the analysed business process are identified from the company's goals. Business questions, corresponding measures and their data sources are identified. No formal algorithm is given to design the conceptual model of a data warehouse using the aforementioned measures and questions. It can be inferred from the model elements that measures are used as attributes of fact tables for the process performance model.

7 Conclusions and Future Work

Our approach offers a method to obtain a data warehouse schema basing on the systematic determination and formalization of analysis needs. The final data warehouse schema can be modified before implementation building on two factors: (1) absence of a particular dimension attribute in a data source and (2) utilization of existing relations between data in data source models to build dimension hierarchies.

In our case study we included only a portion of measurement goals and indicators, but the implementation of the university data warehouse was comprised of considerably bigger number of goals and indicators. Practical results were obtained using the data warehouse and process measurement and monitoring system. They are related to determination of problems during the process monitoring, for example, when it was possible to replan popular lectures, and at the end of the measured process, for example, few part-time students enrolled in courses through internet so there is a possibility to increase the number of process participants.

Some further extensions of our method are possible. The first possible research direction is the comparison of the existing ontologies in the process measurement area with our metamodel to analyze the possibilities of the further development of the proposed method. Another further research direction could be automatic generation of a data warehouse model from OCL query operations. A tool prototype can be implemented for this purpose based on our proposed method.

Acknowledgment

This work has been supported by grant from the European Social Fund.

References

[1] Harrington, J. H., *Business Process Improvement - The breakthrough strategy for total quality, productivity, and competitiveness*, McGraw-Hill, 1991.

[2] Fenton, N.E., Whitty, R., Introduction, *Software Quality Assurance and Measurement, A Worldwide Perspective*, International Thomson Computer Press, 1995, pp. 1-19.

[3] Kueng, P., Hagen, C., Rodel, M., Seifert, S., Business Process Monitoring & Measurement in a Large Bank: Challenges and Selected Approaches, *Proc. of 16th International Workshop DEXA 2005*, 2005, pp. 955-961.

[4] Jarke, M., List,T., Koller, J., The challenge of process data warehousing, *Proc. of 26th International Conference VLDB 2000*, 2000, pp. 473-483.

[5] Kueng, P., Wettstein, T., List, B., A Holistic Process Performance Analysis through a Process Data Warehouse, *Proc. of the AMCIS 2001*, 2001.

[6] List, B., Machaczek, K., Towards a Corporate Performance Measurement System, *Proc. of the ACM symposium SAC '04*, ACM Press, 2004, pp. 1344-1350.

[7] Park, R.E., Goethert, W.G., Florac., W.A., Goal-Driven Software Measurement – A Guidebook. *Technical Report, CMU/SEI-96-HB-002*, Software Engineering Institute, Carnegie Mellon University, 1996.

[8] Basili,V.R., Rombach, H.D, The TAME Project: Towards Improvement-Oriented Software Environments, *IEEE Transactions on Software Engineering*, Vol. 14, No. 6 ,1988, pp. 758-773.

[9] Object Management Group: UML 2.0 Superstructure http://www.omg.org/cgi-bin/doc?formal/05-07-04

[10] List,B., Korherr,B., A UML 2 Profile for Business Process Modelling, *Proc. of ER 2005 Workshops AOIS, BP-UML, CoMoGIS, eCOMO, and QoIS*, Springer, LNCS, Vol. 3770, 2005, pp. 85-96.

[11] Object Management Group: OCL 2.0 Specification http://www.omg.org/cgi-bin/doc?ptc/2005-06-06

[12] Giorgini, P., Rizzi, S., Garzetti, M., Goal-Oriented Requirement Analysis for Data Warehouse Design**,** *Proc. of the 8th ACM Int. Workshop DOLAP*, 2005, pp. 47-56.

[13] Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., Paraboschi, S., Designing Data Marts for Data Warehouses, *ACM Transactions on Software Engineering and Methodology*, Vol. 10, No. 4, 2001, pp. 452-483.

[14] Golfarelli, M., Maio, D., Rizzi, S., The Dimensional Fact Model: a Conceptual Model for Data Warehouses, *International Journal of Cooperative Information Systems*, Vol. 7, No 2&3, 1998, pp. 215-247.