The Processing of business resources using data warehouse based on Hybrid Methodology

Quan Yuan[*]

Hubei University of Technology Engineering and Technology College, Hubei, 430068, China

Received 1 March 2014, www.cmnt.lv

Abstract

In computing, a data warehouse (DW, DWH), or an enterprise data warehouse (EDW), is a system used for reporting and data analysis. In working environment resources are commonly shared between tasks, and sometime multiple resources are necessary to commence a single task, this scenario makes the resource utilization complex. Making it simple to understand this, we classify resources into human resources, because the two types of resources behave differently leaving different impact on process performance. Quality of data produces is evaluated though an empirical study, that is confirming the claim of highly relevant information generation. Data warehouse conceptual design is based on the metaphor of the cube, which can be derived from either requirement-driven or data-driven methodologies. Each methodology has its own advantages. The first allows designers to obtain a conceptual schema very close to the user needs but it may be not supported by the effective data availability. On the contrary, the second ensures a perfect trace ability and consistence with the data sources–in fact, it guarantees the presence of data to be used in analytical processing–but does not preserve from missing business user needs. To face this issue, the necessity emerged in the last years to define hybrid methodologies for conceptual design.

Keywords: Business Resources; Data Warehouse, Hybrid construction

1 Introduction

Integrating data from one or more disparate sources creates a central repository of data, a data warehouse (DW). Data warehouses store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. The data stored in the warehouse is uploaded from the operational systems (such as marketing, sales, etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before it is used in the DW for reporting. In this competitive era, demand of business process improvement increased [1] and it became one of top priority among top ten organizational priorities [2]. Researches on business process improvements established that redesigning of process flow (structure) and resources reassignment are two improvement ways. During the process executions, event traces of process are generated, these traces are acknowledged as process logs [3]. These process logs are widely acknowledged as an important source for posterior analysis. On way of posterior analysis is process mining [4], but information provided by process mining does not yield tempting and attractive analytical relativism in comparison to business intelligent based approaches. It is because data mining always provides similar solutions due to limitations of data availability and unstructured organizations. Business intelligence provides information at different granularity levels according to interest of decision maker though online analytical process (OLAP) systems. These systems can also provide user needed analytical information at strategic, tactical and operational level. One core component of OLAP is data warehouse, a separate repository that is defined as subject-oriented, non-volatile, time variant and integrated collection of data in favor of decision making process [5].

A data warehouse (DW) is a collection of technologies aimed at enabling the decision maker to make better and faster decisions. Data warehouses differ from operational databases in that they are subject oriented, integrated, time variant, nonvolatile, summarized, larger, not normalized, and perform OLAP. The generic data warehouse architecture consists of three layers (data sources, DSA, and primary data warehouse) [6]. Although ETL processes area is very important, it has little research. This is because of its difficulty and lack of formal model for representing ETL activities that map the incoming data from different DSs to be in a suitable format for loading to the target DW or DM [7]. To build a DW we must run the ETL tool which has three tasks: (1) data is extracted from different data sources, (2) propagated to the data staging area where it is transformed and cleansed, and then (3) loaded to the data warehouse. ETL tools are a category of specialized tools with the task of dealing with data warehouse homogeneity, cleaning transforming, and loading problems. This research will try to find a formal representation model for capturing the ETL processes that map the incoming data from different DSs to be in a suitable format for loading to the target DW or DM. Many research projects try to represent the main mapping

[*] ***Corresponding author's*** e-mail: quanyuanhbut@126.com

activities at the conceptual level. Our objective is to propose conceptual model to be used in modeling various ETL processes and cover the limitations of the previous research projects. The proposed model will be used to design ETL scenarios, and document, customize, and simplify the tracing of the mapping between the data source attributes and its corresponding in data warehouse.

Other contribution is to execute analysis queries according to resource hybrid model on process warehouse. It is observed that with useful information a lot of irrelevant information also extracted that make difficult to make decision for an ordinary decision maker. It is due to in adequacy of process warehouse design due to its less customization. One solution is to add hybrid as a fact attribute in fact table to get relevant information. For this versioning operation of PW will be required but, this solution can be very costly due to high versioning cost. As a contribution towards solution of this problem a method is proposed that explains three steps procedure of hybrid evaluations using the resource hybrid model for measuring and analysing hybrid using process warehouse.

The survey of the current methodologies in [8] addresses the necessity of hybrid approaches in multidimensional modelling, defines a unified terminology for multidimensional concepts, and also introduces a set of useful comparison criteria to evaluate methodologies on the basis of their capabilities to adequately represent such multidimensional concepts. An important comparison criterion is the identification of the inputs to be provided to a methodology. As an example, methodologies can work on conceptual or logical schemas, whereas a conceptual schema can be expressed according to the Entity/Relationship (E/R) or the Unified Modeling Language (UML) formalism and a logical schema can be either a relational schema or an XML schema. Other important issues in this research topic are the actual lack of tools supporting automatic multidimensional modelling [8] and the fewness of CASE tools to support automatic design [9]. While methodologies based on a requirement-driven approach suffer from drawbacks related to the comprehension of user needs expressed according to a natural language, in methodologies based on a data-driven approach the adoption of automatic techniques is encouraged by well-structured data sources, and, additionally, by the presence of functional dependencies [10].

The contribution of the paper is the definition of a sequential hybrid methodology that takes into account both the advantages of data modelling, as in the Dimensional Fact Model (DFM) [11], and the strong formalization of user requirements. Such a formalization is represented by UML multidimensional schemas [12] obtained from the framework [13]. These UML multidimensional schemas, due to their high level of standardization and formal representation of multidimensional concepts, can be effectively used for the automation of the design processes. In the paper, we show how a large part of the steps performed in our design of multidimensional schemas can be done automatically.

2 Problem Formulation

Currently, in the available literature, there is no precision regarding DW concept. In some more recent works, it is noticeable that the scope of the term has been increasing, such that it now includes not only data collections, but also support systems for extraction and preparation of data that will compose these collections. It is this broader sense of DW that will be used throughout the present work. Thus, a DW can be interpreted as a system that is designed with the purpose of supporting "efficient (data) extraction, processing and presentation for analytic and decision-making purposes" [14]. With the same scope, Rainardi [15] defined DW as "a system that retrieves and consolidates data periodically from the source systems into a dimensional or normalized data store".

In the current scenario, there are opportunities for the use of all these approaches, depending on the goal to be reached through construction of the DW. Many organizations are closer to the idea developed by Kimball, because DWs often arise from the union of several dimensional databases developed by departments in independent efforts. Additionally, some database writers say that they prefer a fully dimensional approach because this model is simpler for end users to understand and interpret. However, this claim has been questioned by some researchers [16]. The use of an approach closer to that proposed by Inmon has been advocated because of the ease of maintenance of a central and historical database when it is normalized. Data consistency is more easily ensured in this manner. Small data marts, always fed by the NDS, can then be constructed as needed. In the present work, one of our major concerns was to integrate data coming from several heterogeneous blood center systems, while bearing in mind their quality during the integration process. Therefore, we decided to adopt the approach proposed by Inmon, which facilitates data maintenance and consistency. Just as important as the DW databases are the components that manipulate the data in order to populate these databases. One very common example of components for this purpose is the ETL (Extract, Transform and Load) tools [17], which have the primary function of facilitating the following processes: extracting data from various systems; transforming, validating and correcting extracted data as necessary; adapting them to the needs of the DW; and loading the data into analytical or normalized data repositories. There are many uses for such tools, even outside of the DW context.

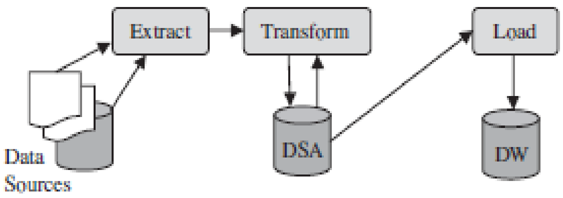

Figure 1. Bus matrix for sewer pipe data warehouse..

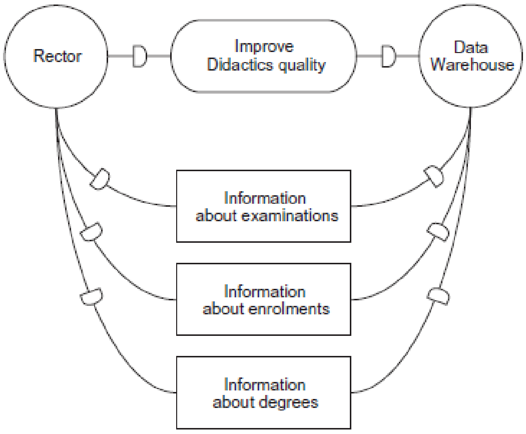
Rector
Improve Didactics quality
Data Warehouse
Information about examinations
Information about enrolments
Information about degrees

Figure 2. The metamodel for the logical entities of the ETL environment.

3 The Proposed Analysis Framework

One of the critical processes in designing a data warehouse is to determine what data should be extracted and loaded into the data warehouse from the various data sources. Rujirayanyong and Shi [18] introduced two strategies for the identification of project data sources: the need-based approach and the availability-based approach. In the availability-based approach, any data about operational systems that is currently available is selected and uploaded to the data warehouse. The need-based approach, however, investigates the potential need for future analysis of data by considering the business nature of the system. In our study, the need-based approach was utilized to develop an optimal data warehouse to improve existing sewer management systems. This approach was chosen because, in sewer management, needs were relatively clear whereas the available data were scarce.

The next step is the structural design process. Usually, this process is composed of four steps, including selection of business processes, declaration of granularity, determination of associated dimensions, and identification of the facts [19]. The selection of business processes requires understanding of business requirements and associated data for the systems. Granularity refers to the required level of detail for information stored in a data warehouse [20]. In general, the level of detail is subject to the interests of users and to the amount of relevant data collected. Once the granularity of the fact table is determined, the related dimensions are reasonably obtained based on the nature of the fact table. Facts reflect the objective of the data warehouse and, accordingly, should be composed of data useful for decision makers.

In our study, implementation of the proposed data warehouse followed the basic procedures discussed above. In the current practice of sewer management, the main business processes are installation, inspection, and renewal activities. These processes, along with the corresponding dimensions, can be represented by the data warehouse bus matrix (DWB), which is shown in Fig. 1. The data warehouse bus structure was first introduced to ensure consistency in data warehouse design. In data warehouse architecture, the matrix row represents the relevant business processes that users elect to monitor.

On the other hand, matrix columns represent conformed dimensions against which business processes are measured. As shown in Fig. 1, the installation process is measured against four common dimensions, including pipes, materials, basin codes, and dates, whereas the renewal process is measured using five common dimensions. The successful construction of a bus structure for sewer pipe management requires several aspects of effective structural design. First, the bus structure should include all pertinent business processes. The list of processes need not be limited to the three shown in Fig. 1; depending on the specific condition of the particular infrastructure, other processes, such as policy making and maintenance, may be included as needed. Second, the associated (common) dimensions should be identified and consistently standardized to avoid excessive work on the collection of data. Finally, the architecture of the bus structure should provide a foundation for useful interaction among the data marts or organizations. As shown in Fig. 1, well-developed common dimensions, such as pipes, materials, and basin codes, are all shared by the three main business processes, which allows for smooth interaction and integration of sewer management processes.

4 Data Warehouse based Processing Model

In order to maintain the best features of either the presented methodologies, we utilize a hybrid methodology that covers both requirement analysis and conceptual design. The aim is to produce multidimensional schemas (or fact schemas) that capture user needs but also allow designers to execute data modelling activity.

The main idea is to consider reconciled UML schemas, obtained from the framework, and the UML multidimensional modelling, as the input to an advanced data modelling step. So, our hybrid methodology is based on two multidimensional models: (1) the UML for data warehouse to represent requirement-derived multidimensional schemas, and (2) the extended DFM to represent a tree-based view of the data sources. This view, namely the attribute tree, allows the designer to easily manipulate the structure of the underlying data sources. In fact, according to this model, the modification of functional dependencies in data sources corresponds to intuitive operations on the tree, such as adding and/or removing nodes.

The extended DFM is a model defined to improve some features of DFM. In fact, the algorithm used in the DFM is not able to manage many-to-many relationships and it stops the generation of the branch of attribute tree when this kind of relationship is encountered. Therefore, in this context, it is not possible to use the DFM for two main reasons: (a) the algorithm runs only on relational (or E/R) schemas of the data sources, and (b) it cannot establish many-to-many relationships among entities in automatic way.

On the other hand, the extended DFM is equipped with constructs to represent many-to-many relationships and a logical program able to build the attribute tree from any relational schema containing binary relationships.

In our hybrid methodology, the advanced data modelling is based on the extended DFM, while the multidimensional modelling follows the conventional hybrid methodology to produce reconciled UML schemas. In order to use reconciled UML schemas as input to the data modelling activity, we cannot use the existing logical program but we must face with the schema translation issue. That is to say, given a model M2 and a schema S1 over a model M1, to find a schema S2 over M2 which is equivalent to S1. In this context, M1 is the UML for data warehouse and M2 is the extended DFM. Although these models quite differ in their terminology and graphical constructs, they allow to represent the same multidimensional elements, in the sense that each element of M1 has its counterpart in M2 and vice versa.

5 Experiment and Results

The proposed system for sewer management includes four decision supporting modules: import, decision/estimation, report, and export modules. As developed for this study, the decision support system was equipped with a user-friendly graphical user interface (GUI) in the Microsoft Excel environment for effective handling of infrastructure information. SQLite (structural query language) version 3.7.3 (Fig. 2) was used to implement the import and export modules of the system. Once the “import” button on the GUI was clicked, the infrastructure data stored in the data warehouse was transferred to the Excel environment. Subsequently, the “decision/estimation” module appropriately determined the most appropriate options for inspection and renewal methods without any additional support from experts. At the same time, the “decision/estimation” module computed approximate costs for inspection and renewal of each pipeline. Use of the “report” module enabled the system, with the help of the Excel pivoting tool, to provide seven different types of managerial reports for various users with diverse levels of detail. For example, a range of users, such as policy makers, project schedulers, and project engineers, could use the system at varying management levels, such as city, sewer basin, and sewer shed. Finally, estimated costs and selected methods of inspection and renewal were exported to the data warehouse through the “export” module. Although the SQL-based data warehouse was able to accommodate and visualize valuable information, it provided limited applicability for data analysis due to compatibility issues with other software and prevalence issues for users. Thus, Microsoft Excel software was adopted in this study to overcome these shortcomings. Since Excel is one of the most widely used software for data analysis, it provides an extremely flexible environment for sewer management. Once data was exported to Excel, users could analyze the data from various points of view with built-in functions. Moreover, in the proposed decision supporting modules, the linkage between the data warehouse and Excel was easily accomplished using Visual Basic for Application (VBA) programming. As a result, the decision supporting modules provided enhanced flexibility for data analysis.

In dealing with DW systems in which the data come from heterogeneous systems, “we cannot boil the ocean”. It is a hard task to understand all the details of the available data that come from multiple sources at once. To facilitate development of such a system, it is important to focus on entities that are common among all the sources. Through these entities, a starting point for building a conceptual model is reached and, from this, for building the DW. After the first version has been developed, it can be used to populate the databases and generate some views and reports. The analyses on these views will then be very useful in verifying whether any errors exist in the concepts that have been modeled so far. It will then be possible to develop a new version of the conceptual model, containing corrections for detected errors. After completing enough cycles of analyses and development for a reasonably stable version of the DW to be achieved, expansion of the conceptual model can be envisaged, such that it will cover more entities or attributes, thus restarting the cycle of analyses and corrections. All these changes will be easily accomplished if modular architecture with high flexibility and loose coupling is used. Combined use of the practices that we have discussed here (modular architecture, conceptual modeling and data analyses) from a cyclic and interactive perspective allowed us to reduce the complexity of DW development. In other words, by following these good practices, we quickly and efficiently achieved a DW that contained the most interesting entities stored and with the possibility of generating managerial views and reports from these data. We were therefore able to solve the main cause of DW development failure: the delays in delivering a functional system.

6 Conclusion

The requirement-driven approach is the only method able to fully capture user needs. In this context, the i_ methodology is one of the current approaches to reach this goal. The result of the design is a conceptual model expressing the multidimensional concepts according to the UML standard. An extension of this methodology consists of the reconciliation of UML schemas with the data sources, obtaining a multidimensional model that both captures user needs and, furthermore, agrees with data sources. We used this model as the starting point for a novel approach. We consider a reconciled UML schema as a surrogate of the conceptual schema representing data sources and, then, use it in a data-driven methodology, as the DFM. Following this guideline, the designer works on a conceptual schema containing all and only the useful attributes and does not risk the danger to keep the unnecessary ones, determining a decay of the data warehouse quality, as our metrics evaluation reports. Moreover, the designer is not forced to correctly map multidimensional concepts to entities in data sources. This is a relevant feature for, in a very complex E/R schema, there can be serious difficulties in choosing those to be considered as possible fact tables. On the contrary, in a reconciled UML schema, multidimensional concepts have already been identified and mapped to data source using QVT.

Acknowledgement

The work was funded by Scientific Research Foundation on the Introduction of High-level Talents of Hubei University of Technology Engineering and Technology College.

References

[1] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211.

[2] Burattin, A., & Sperduti, A. (2011). PLG: A framework for the generation of business process models and their execution logs. In Business Process Management Workshops (pp. 214–219). Springer.

[3] Conner, M., & Armitage, C. J. (1998). Extending the theory of planned behavior: A review and avenues for further research. Journal of Applied Social Psychology, 28(15), 1429–1464.

[4] Glasser, W. (1998). Choice theory: A new psychology of personal freedom. HarperCollinsPublishers.

[5] Goldsmith, J., & Junker, U. (2009). Hybrid handling for artificial intelligence. AI Magazine, 29(4), 9.

[6] Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the hybrid reversal phenomena. Psychological Review. doi:10.1037/0033-295X.94.2.236

[7] Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., & Shan, M. C. (2004). Business Process Intelligence. Computers in Industry, 53, 321–343. doi:10.1016/j.compind.2003.10.007

[8] Heravizadeh, M., Mendling, J., & Rosemann, M. (2009). Dimensions of business processes quality (QoBP). In Business Process Management Workshops (pp. 80–91). Springer.

[9] Hong, S., Katerattanakul, P., Hong, S.-K., & Cao, Q. (2006). Usage and perceived impact of data warehouses: A study in Korean financial companies. International Journal of Information Technology & Decision Making, 5(02), 297–315.

[10] Huang, Z., Lu, X., & Duan, H. (2012). Resource behavior measure and application in business process management. Expert Systems with Applications, 39(7), 6458–6468.

[11] Kumar, A., Dijkman, R., & Song, M. (2013). Optimal resource assignment in workflows for maximizing cooperation. In Business Process Management (pp. 235–250). Springer.

[12] Lee, K. M. (2004). Adaptive resource scheduling for workflows considering competence and hybrid. In Knowledge-Based Intelligent Information and Engineering Systems (pp. 723–730). Springer.

[13] List, B., Bruckner, R. M., Machaczek, K., & Schiefer, J. (2002). A comparison of data warehouse development methodologies case study of the process warehouse. In Database and Expert Systems Applications (pp. 203–215).

[14] Macris, A., Papadimitriou, E., & Vassilacopoulos, G. (2008). An ontology-based competency model for workflow activity assignment policies. Journal of Knowledge Management, 12(6), 72–88.

[15] Ouyang, C., Wynn, M. T., Fidge, C., ter Hofstede, A. H. M., & Kuhr, J.-C. (2010). Modelling complex resource requirements in business process management systems. ACIS 2010 Proceedings.

[16] Richter, M. K. (1966). Revealed hybrid theory. Econometrica: Journal of the Econometric Society, 635–645.

[17] Sen, A. (1973). Behaviour and the concept of hybrid. Economica, 241–259.

[18] Sen, A. K. (1971). Choice functions and revealed hybrid. The Review of Economic Studies, 307–317.

[19] Shahzad, K., & Sohail, A. (2009). A systematic approach for transformation of ER schema to dimensional schema. In Proceedings of the 7th International Conference on Frontiers of Information Technology (FIT '09) (p. 6). doi:10.1145/1838002.1838060

[20] Shahzad, M. K. (2012). Improving Business Processes using Process-oriented Data Warehouse.

Authors

< Quan Yuan >, 1982. 10, Wuhan, Hubei Province, China

Current position, grades: Associate Professor
University studies: E-commerce
Scientific interest: E-commerce, Logistics and Supply Chain Management
Publications: 20 papers published in the international or national journals
Experience: She is an associate professor in Hubei University of Technology Engineering and Technology College, China. Her research interests include E-commerce, logistics and supply chain management.