VOL. 10, NO. 3, FEBRUARY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

QUALITY ORIENTED FOR PHYSICAL DESIGN DATA WAREHOUSE

Munawar, Naomie Salim and Roliana Ibrahim
Department of Information System, Universiti Teknologi Malaysia, Malaysia
E-Mail: an_moenawar@yahoo.com

ABSTRACT

To construct a data warehouse (DW) as a collection of data marts (DMs) all at once in a single project is very difficult. There is a need to determine the first DM that should be constructed. After two or more DMs are constructed, the next problem is integrating all DMs into the enterprise DW. This paper is adaptation of bus matrix, quality function deployment (QFD) matrix, political factors and DM integration technique to be a single framework in order to determine the priority of DMs that should be constructed and how to integrate the entire DMs into the enterprise DW.

Key words: data mart construction priority, data mart design, data marts integration, physical design.

INTRODUCTION

Data warehouse (DW) development phase ends in physical design with respect to architecture design of a DW and its related to data marts (DMs). Including this phase is physical optimization techniques such as materialized view and drill-across query.

Recently, common practice for building a DW is bottom-up design. In bottom-up design, DW is a collection of integrated DMs (Kimball and Ross, 2002), where each DM is dedicated to study of single subject in the enterprise. The most important reason to choose this approach is faster and cheaper, while building the whole DW as a single project is very expensive and requires long time (Diamantini and Potena, 2012).

The first step to construct a DM is identifying all the facts to be placed in the DW (Golfarelli, Maio and Rizzi, 1998). The collection of facts required by the enterprise will evolve over time, and can be in years to be completed. Kimball and Caserta (2004) recommends to construct the first DM that represents the minimum effort and risk. The first DM can become a foundation on which others will be built and can be reused by other DMs. But, there is no detailed guidance to determine the priority of DM to be built. McFadyan and Chan (2001) suggested using quality function deployment (QFD) matrix as a useful tool to justify and support the DM construction strategy. However, political factors are not considered explicitly within this matrix. DW projects are always potentially political because they influence the work practice of highly autonomous and powerful user communities in an enterprise (Demarest, 1997) and contribute in most common reasons for DW project failure (Watson, et. al, 1999)

Although the DM idea offers the opportunity to build a corporate DW from the bottom-up, the benefits of DMs can easiliy be outweighed and cross-functional analysis is not possible (Houari and Far, 2004). Cross-functional analysis makes information more easily and broadly accessible. As a result, business can gain competitive advantage and realize greater business value through the integration of DMs.

Data overlapping betwen two or more DM often happens in multiple DMs. If done properly, it will be very beneficial. Otherwise, it can cause severe problem in architecture (Sumathi and Sivanandam, 2006)

DMs (subset of DW) are conformed by following a standard set of attribute declarations called a DW bus (Kimball and Ross, 2002). Data quality (DQ) has to be incorporated in DM design in order to avoid loss of data, inadequate dimension hierarchy, incorrect data aggregation and violating the additivity of facts with respect to dimensions (Gamal, Bastawissy, and Galal-Edeen, 2011). Lack of proper DM schema could lead to misleading the decision makers with incorrect information. Defects of DQ will eventually lead to failure in providing accurate business information. Each of these DQ dimensions has a great influence on the overall quality of DW. By incorporating quality function, we can add value to our technique, and finally DW project failure can be minimized. Therefore, there is a need to incorporate DQ in DM design phase

This paper is our next step in integrating DQ into the whole phase of DW development. After determining the framework for DW development (Munawar, Salim and Ibrahim, 2011a), we continue with comprehensive study of quality dimensions for DW development (Munawar, Salim and Ibrahim, 2012). After that, describing the detail requirements analysis and conceptual design in incorporating DQ is become our concern (Munawar, Salim and Ibrahim, 2011a and 2014a). Next, integrating DQ into the logical design (Munawar, Salim and Ibrahim, 2011a and 2014b) and finally, in this paper we try to elaborate bus matrix, QFD matrix and DQ dimensions to ensure the quality of DW construction.

The rest of this paper is organized as follows. Next section presents the data mart design approaches. After that proposed framework to determine the priority in constructing the DM is explained, and then followed by implementating the proposed framework for academic affairs in a private university. Finally, this paper ends with conclusion and future work.

DATA MART DESIGN APPROACHES

There are two major approaches for DM design. The followings are detail explanation about these approaches.

Dependent data mart

A dependent DM is a subset of a DW (Inmon, 1996), focused on single business process of enterprise. It is top-down approach where DM is created by DW. Assumption that once DW is constructed, automatically the DM will become a subset of a DW is flawed. In the top- down approach, DM that was generated subject to departments or users feedback. Therefore, gradual evolution is needed for the DW and the DM.

Independent data mart

Independent DM is a bottom-up approach, where DM can be created firstly and then a DW can be generated using multiple DM (Kimball, 1996). Therefore, there is a need for DM integration. DM integration is the process for emerging data at different DM and providing a single view of this data. There are three approaches for data integration in data mart (Chhabra and Pahwa, 2014).

- Integration with dimensions sharing
  Multiple DMs can be joined ove common dimensions (Kimball, 1996). Having shareable dimensions gives us the change to integrate the DM through the same dimensions in both of DM
- Integration with dimensions compatibility
  Two dimensions in different DM can be said compatible when their common information is consistent or their contents can be combined in a meaningful way (Chhabra and Pahwa, 2014; Cabibo and Torlone, 2004)
- Integration with generalization
  It still possible to integrate the DM via drill-across query eventhough the dimensions are not same (Abelo, Samos, and Saltor, 2002). Dimensions in the different DM can be connected by generalization.

PROPOSED FRAMEWORK

Despite the existence of several methods and tools addressing the problems related to the implementation of the DMs, little attention has instead been paid to the design of the DM. The basic principle underlying the proposed approached is that design of DMs should be driven by the business needs that each DM is expected to address.

Bus matrix maps the business processes to the entities or objects that participate in these processes. The bus matrix is essentially our enterprise dimensional data architecture. For each business process (row), we can see exactly which dimensions (columns) we need to implement. And for each dimension, we can see which business processes it must support. A data warehouse quality function deployment (DWQFD) is useful tool for planning the incremental/ evolutionary DW construction (McFadyan and Chan, 2001). However, political factors and quality factors are not considered explicitly. DW project will become politicized in the following condition (Demarest, 1997): (1) strong but poorly-defined sense of urgency that can not be clearly linked to changes in the firm's market position or financial health of firms (2) cost justification without clear indications of the target for that justification (3) economic impact to the firms (Payton and Zahay, 2005). Therefore, organizational politics is integral part of DW projects

Many independent DMs will be developed over years in the DW project. The next problem in building enterprise DW is integrating heterogenous DMs. The problem of integrating heterogenous DMs is identifying competible dimensions. *Conformed* (identical in semantics, structure and data) is required by Kimball for integrating dimension table (Kimball and Ross, 2002). Without conformity, there is no possibility to combine or perform meaningful aggregation of measures across different data marts due to the loss of data resulting from the join between these dimensions tables (Cabibbo and Torlone, 2004). This is clearly restrictive and difficult to achieve when integrating heterogeneous data marts (Torlone, 2008).

The success of DMs integration can be reflected by the same condition on the dimensions and aggregation hierarchies for the DW between before and after integration.

However, in many practical applications, especially in independent DMs, the aggregation hierarchy for dimensions are not always available (Riazati, Thom and Zhang, 2010).

Dimensions integration in the DW can significantly improve the quality of decision making process. The consolidation of DMs via dimensions integration produces a single, consistent, integrated and accurate view of the data within the organization. As a result, business can gain cost savings, cost avoidance, better return on IT investments, better address information security & regulatory compliance and finally improve decision making that drives competitive differentiation.

As an integral part of DW development, physical design plays an important role in addressing performance of the DW. In relation with DM integration and data quality, physical design deals with drill-across query and materialized view.

Combining and correlating data from multiple data marts and to perform value chain analysis are the main goal of *drill-across* queries (Kimball and Ross, 2002). Common dimensions are prerequisite for joining different DMs via drill-across queries. Therefore, it is important to have compatible dimensions and facts in order to see consistently at data across DMs and to combine and correlate such data, e.g., to perform value chain analysis.

Materialized view can be used to improve data access time by precomputing intermediary result. The cube with lower granularity can be materialized in order to do not loose information. The most frequent cubes that need to be improved or complex queries are also reasonable to be materialized. If there is no feasibility to materialize all possible cubes, then, the decision is a trade-off between storage and performance.

Here, we combine bus matrix, DWQFD, political factors, and data quality dimensions to construct physical design of the DW. The physical design interprets the DW architecture as a network of data stores, aiming at the quality factors of reliability and performance in the presence of very large amounts of slowly changing data. *Slowly changing dimensions* (SCDs) is a well-known concept in dimensional modeling of DWs that accounts for the *potential volatility* of dimension members. SCDs preserve a history of evolving dimension instances, and thus allow tracing and reconstructing the correct dimensional context of all measures in the cube over time. The general purpose of SCDs is to preserve the referential integrity between facts and dimensions in hyper-cubes while supporting changes in dimension members (Kimbal, 2002). Proposed framework can be depicted in the following Figure-1.

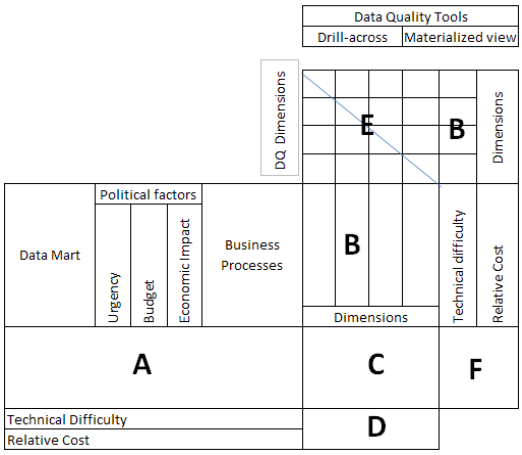

Figure-1. Proposed framework.

Figure 1 shows the following information:

- The data marts to be built and their political factor that influence the decision to develop a DW (Area A)
- Required dimensions to support business process. (Area B)
- Mapping the required dimensions in every data mart. (Area C)
- The technical difficulty and relative cost of constructing a dimension (Area D)
- The relationship between dimensions (Area E). Some dimensions will be viewed by others; some will be sourced from the same legacy tables.
- Justification to rank the DM construction strategy. (Area F)

EVIDENCE PRACTICE

A private university intends to develop a DW to support its decision maker in relation with academic affair. Using the template in Figure-1, we tried to determine the first DM that should be constructed in academic affair. Figure-2 shows the detail explanation about the implementation of this framework.

Consider Figure-2 regarding the following information:

- The top matrix shows the relations between the dimensions in academic environment.
- Time is the easiest dimension from technical aspect, whereas the student may be the most difficult dimension. The student dimension may be sourced from many tables in many databases that evolved over a period of decades
- The relative cost of building dimensions will be corresponded to many elements such as the technical difficulty, the number of related table, joining or matching complexity, and the transformations complexity.

The information in this matrix can be used to justify and support the DM construction strategy. The higher total political factor and the lower total cost can be used to rangk the data mart that should be constructed. Even though the cost and technical difficulty of enrollment DM (total 97) is higher than Quality Assurance DM (total 34), however the political factor for enrollment DM (total 20) is higher than Quality Assurance DM (total 9). Therefore, combination between relative cost, technical difficulty and political factor can be said that enrollment DM (total 125 from 28+97) is in rating number one and be the first DM to be built. The second DM should be Class Administration DM. It represents the lower technical difficulty and its relative cost (total 124) then Registration Semester DM (total 142). However the political factor of Class Administration DM (total 29) is higher than Registration Semester DM (total 26). Therefore, Class Administration DM can be said as the second DM to be constructed.

Figure 2 also shows that some dimensions can be shared: time, department, faculty, basis, and student. All DMs in academic environment in case study can be integrated all together using these five dimensions.

More clearer mapping between all dimensions in the entire DMS, give us easiness of integrating the whole DMs into the enterprise DW through dimensions sharing, dimension compatibility and generalization of dimensions. Dimensions sharing can be read in the DWQFD matrix from the analysis, dimensions compatibility can be seen from the same attribute in the dimensions, whereas generalization (if necessary) can be used to connect uncompatible different dimensions.

Using DWQFD matrix as depicted in Figure-2 gives us the advantages as follows:

- Priority of the DMs that should be constructed can be determined easily
- Shareable dimensions can be mapped easily
- DWQFD matrix can be treated as a blue print for DW construction. Therefore, evolution of the DW in different time, changing the designer and tools will give minimal effect in DM integration.

Data Quality Tool																
Drill-across Queries									Materialized View							
Data Quality Dimensions	Business Quality	Information Quality	Technical Quality													Dimensions
				X					X						Time	
					X										Department	
						X									Faculty	
														X	Financial	
							X		X						City	
									X						Basis	
					X										School of Origin	
															Event	
					X	X									Student	
															Payment Scheme	
												X			Curriculum	
															Employee	
											X			X	Subject	
															Room	
				X								X			Lecturer	

Data Mart	Political Factors: Total	Urgency	Cost	Economic Impact	Business Processes	Time	Department	Faculty	Financial	City	Basis	School of Origin	Event	Student	Payment Scheme	Curriculum	Employee	Subject	Room	Lecturer	Technical difficulty	Relative Cost	Total
						Dimensions																	
Enrollment	28	10	8	10	Admision	X	X	X	X	X	X	X	X	X	X	X	X				5	15	
					Test	X	X	X													10	15	
					Registration	X			X					X	X						7	15	
					Reward Calculation	X	X	X	X		X	X		X	X		X				10	20	
																					32	**65**	**97**
Registration Semester	26	5	15	6	Course Offer		X	X			X					X		X			4	8	
					Scheduling		X	X			X					X		X	X		12	25	
					Registration Semester	X					X			X	X			X	X		4	7	
					Students Billing	X	X	X	X		X			X	X						5	10	
					Students Payment	X	X	X	X		X			X	X						5	10	
					Class Allocation	X	X	X			X			X	X	X		X	X	X	12	25	
					Lecturer Allocation	X	X	X			X					X		X	X	X	5	10	
																					47	**95**	**142**
Class Administration	29	4	20	5	Time Sheet	X	X	X			X			X				X	X	X	7	10	
					Teaching Invoicing	X	X	X	X		X							X		X	10	15	
					Teaching Payment	X	X	X	X		X			X				X		X	12	25	
					Exam Scheduling	X	X	X			X							X	X	X	8	17	
					Exam Grading		X	X			X			X				X		X	7	13	
																					44	80	**124**
Quality Assurance	9	1	5	3	Transcript Generation		X	X			X			X		X		X		X	5	8	
					Outstanding Billing Process		X	X	X		X			X	X						9	20	
																					14	**28**	**34**
Technical Difficulty						1	3	2	7	2	3	2	3	10			3	5	4	4			
Relative Cost						1	3	2	10	2	5	2	4	20			3	15	10	8			

Figure-2. DWQFD matrix for academic affair in a Private University.

CONCLUSIONS AND FUTURE WORK

Develop a data warehouse as data marts collection is an iterative process. Building all data mart at once in a single project is very difficult. There is a need to define the first data mart to be constructed. Justification to the data marts construction strategy can be done by using our proposed framework.

Based on simulation in academic environment, priority in DM construction can be justified. Using the same technique, dimensions sharing can also be explored in order to integrate all DMs that will be constructed. Mapping all dimensions into the entire DMs, gives us easiness of integrating the whole DMs into the enterprise DW.

To be a single framework, physical design of DW should be treated as an integral part of the DW development. Therefore, our future work is synthesizing our proposed framework with our previous works in requirements analysis, conceptual design, and logical design to be a single framework for the DW development.

REFERENCES

Abello A., Samos J. and Saltor F. (2002). On Relationships Offering New Drill Across Possibilities. In Int. Workshop on Data Warehousing and OLAP (DOLAP2002), ACM.

Cabibbo L. and Torlone R. (2004): Dimension compatibility for data mart integration. Proceedings of the 12th Italian Symposium on Advanced Database Systems, Cagliari, Italy, pp. 6-17.

Chhabra R. and Pahwa P. (2014). Data Mart Designing and Integration Approaches. Internation Journal of Computer Science and Mobile Computing. IJCSMC, Vol 3, Issue 4, pp. 74-79. ISSN 2320-088X.

Demarest M. (1997). The Politics of Data Warehousing.

Dimantini C. and Potena D. (2012). Data Mart Integration at Measure Level. M. De Marco et al. (eds.), Information Systems: Crossroads for Organization, Management, Accounting and Engineering, DOI 10.1007/978-3-7908-2789-7_15, Springer-Verlag Berlin Heidelberg.

Gamal N., Bastawissy A. and Galal-Edeen G. (2011). Towards a Data Warehouse Testing Framework. Ninth International Conference on ICT and Knowledge Engineering. IEEE.

Golfarelli M., Maio D. and Rizzi S. (1998). Conceptual Design of data warehouses from E/R schemes. Proceedings of the Hawaii International Conference on System Sciences.

Houari N. and Far B., H. (2004). An Intelligent Project Lifecycle Data Mart-Based Decision Support System. CCECE 2004 CCGEI 2004.

Inmon W. H. (1996). Building the Data warehouse. John Wiley & Sons, Second Edition.

Kimball R. and Ross M. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons, Second edition.

Kimball R. and Caserta J. (2004). The Datawarehouse ETL Toolkit: Practical techniques for Extracting, Cleansing, Conforming and Delivering Data. Wiley Publishing, Inc. IN 46256. ISBN : 0-764-56757-8.

Kimball R. (1996). The data warehouse tool kit , John wiley & sons.

McFadyan R. and Chan F. (2001). QFD Matrix for Incremental Construction of a Warehouse via Data Marts. H. Balsters, B. de Brock, and S. Conrad (Eds): FoMLaDO/ DEMM 2000. LNCS 2065, pp. 133-141, 2001. Springer-Verlag Berlin Heidelberg.

Munawar Salim N. and Ibrahim R. (2011a). Toward Data Quality Integration into the Data Warehouse Development. Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing. 978-0-7695-4612-4/11 © IEEE Computer Sociaty. DOI 10.1109/DASC.2011.194.

Munawar Salim N., Ibrahim R. (2011b) Toward Data Warehouse Quality through Integrated Requirements Analysis. In: ICACSIS 2011 ISBN: 978-979-1421-11-9.

Munawar Salim N. and Ibrahim R. (2012). Comparative Study of Quality Dimensions for Data Warehouse Development : A Survey. A. Ell Hassanien et al. (Eds.): AMLTA 2012, CCIS 322, pp. 465–473, 2012. © Springer-Verlag Berlin Heidelberg.

Munawar Salim N. and Ibrahim R. (2014a). Quality-Based Framework for Requirement Analysis in Data Warehouse. International Conference on Advance Informatics : Concepts, Theory and Application. ISBN: 978-1-4799-6984-5.

Munawar Salim N. and Ibrahim R. (2014b). Quality Oriented Logical Design for Data warehouse Development. International Conference on Advances in Computer and Electronics Technology (ACET) ISBN: 978-1-63248-024-8.

.

Payton F.C. and Zahay D. (2005). Why does not Marketing Use the Corporate Data Warehouse? The role of trust and quality in adoption of data-warehousing technology for CRM applications. Journal of Business & Industrial Marketing © Emerald Group Publishing Limited [ISSN 0885-8624] [DOI 10.1108/08858620510603918].

Sumathi S. And Sivanandam S.N. (2006). Data Marts and Data Warehouse: Information Architecture for the Millennium, Studies in Computational Intelligence (SCI) 29, 75–150 Springer-Verlag Berlin Heidelberg.

Torlone R. (2008). Two approaches to the integration of heterogeneous data warehouses. Distrib. Parallel Databases 23(1), 69–97.

Riazati D., Thom J.A. and Zhang X. (2010). Inferring Aggregation Hierarchies for Integration of Data Marts (2010). P. Garc´ıa Bringas et al. (Eds.): DEXA 2010, Part II, LNCS 6262, pp. 96–110, 2010. Springer-Verlag Berlin Heidelberg.

Watson H. J., Gerard J. G., Gonzalez L. E., Haywood M. E., and Fenton D. (1999). Data Ware-housing Failures: Case Studies and Findings, Journal of Data Warehousing (4:1), 1999, pp. 44-55.