A Step towards Centralized Data Warehousing Process: A Quality Aware Data Warehouse Architecture

Maqbool-uddin-Shaikh
Comsats Institute of Information Technology Islamabad
maqboolshaikh@comsats.edu.pk

Sarah Yaqoob
Comsats Institute of Information Technology Islamabad
sarah.yaqoob@gmail.com

ABSTRACT

Data Warehouse is becoming more and more popular by providing business an edge over their competitors. DW is developed by using DWP methodology. Currently, there are large numbers of methodologies followed in market. The reason for this is the lack of any centralized attempts of creating platform-independent DWP standards. For developing a centralized DWP it is important to understand the existing methodologies. The authors have done a comparison of five well known methodologies and highlighted the similarities between the processes they used. A new DW architecture has been proposed by integrating those similar processes. This new architecture concentrates mainly on the quality of the DW, as it is one of the critical aspects of DW. Quality is being introduced by defining 3 new components named as quality control, DW monitor, DW Integration Change Management. Evolution of metadata and DW repository are the most important tasks of the quality management.

Keywords: Data warehouse (DW), data warehousing process (DWP), data warehouse architecture (DWA).

1. Introduction

A Data Warehouse (DW) is a collection of technologies aimed to enabling the knowledge worker (executive, manager, analyst, etc) to make better and faster decisions. It is expected to present the right information in the right place at the right time with the right cost in order to support the right decision making and planning. DW has become an essential component of modern decision support systems.

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process" [1].

Implementation of DW is generally based on a data warehousing process (DWP) methodology [2]. According to latest research there are more then 30 methodologies followed in the data warehouse market. Each of the available methodology is somehow different from the other, having its own set of data warehousing process (DWP) tasks [3]. They are characterized as the vender-specific methodologies. There is a lack of vender-neutral and plat-form independent methodology in the data warehouse market.

The objective of this paper is to emphasize the need of a centralized data warehousing process (DWP) methodology and comparing five of the renowned methodologies used by different core-technology vendors and information modeling companies. Based on this comparison an attempt is being made to make a centralized DWP methodology by integrating some of the sub tasks of DWP with the data warehouse architecture (DWA) by proposing a quality aware DWA.

This paper proposes a quality aware data warehouse architecture (DWA) and quality management framework. The main contributions include an extension of the standard DWA used in the literature. Author's goal is to enable a computationally tractable yet very rich quality analysis, and a quality-driven design process.

The rest of the paper is organized as follows. In section 2 authors describe the motivation for this research; section 3 explains the data warehouse domain. Section 4 contains the comparison of five methodologies, quality aware DWA is explained in section 5 and the paper ends with conclusions in section 6.

2. Motivation

Organization's gain competitive advantage through system that automates business processes by offering more reliable and efficient system to the customers.

These are called as Online Transaction Processing (OLTP) systems or operational systems. With the advent of such systems the organizations resulted in the growing amount of operational data. Organizations then focus on the ways to use this data for corporate decision making.

OLTP systems are not suitable for decision-support quires because every system is designed with a different propose to fulfill different set of requirements. OLTP systems were designed to maximize transaction processing capabilities while decision making requires response to ad-hoc quires. These quires are usually complex and involve analytics such as aggregation, drill-down, and slicing/dicing of data.

These quires come under the realm of Online Analytical processing (OLAP), which is defined as "the dynamic synthesis, analysis, and consolidation of large volumes of multi-dimensional data" [1]. Decision Support System (DSS), Executive Information System (EIS), Geographic Information System (GIS) and several other analyses and reporting tools are expected to ask such type of ad-hoc quires for their decision making and functioning from the DW.

Authors in [4, 5] highlighted the need of a centralized DW development process by making an ontological model of the existing DWP methodologies. Research has shown that DW has been cited as the highest priority post-millennium project of more than half of IT executives [4]. A recent study conducted by the Meta Group found that 95% of the companies surveyed intended to build a DW.

Interaction with DW tool vendors, DW application developers and administrators has shown that the standard framework used in the DW literature is insufficient to capture in particular the business role of DW. The construction of DW is a major investment made to satisfy some business goal of the enterprise. Quality model and DW design should reflect this business goal as well as its subsequent evolution over time. Several attempts have been made to enhance the quality of DW by concentrating on some of the quality goals [6].

3. Data Warehouse Realm

While looking into the realm of DW two things are very important, DWA and DWP. The DWA provides the tools that are important for creating a DW within an organization. The DWP deals with the processes that are used to create the DW. The details of these concepts are as follows;

3.1 Data Warehouse Architecture

Architecture is a blueprint that allows communication, planning, maintenance, learning, and reuse. It includes different areas such as data design, technical design, and hardware and software infrastructure design [4].

Several architectural designs for DW are available. Some of the common designs are: data mart architecture, centralized DWA, and centralized data warehouse with dependent data marts architecture (Figure 1) [4].

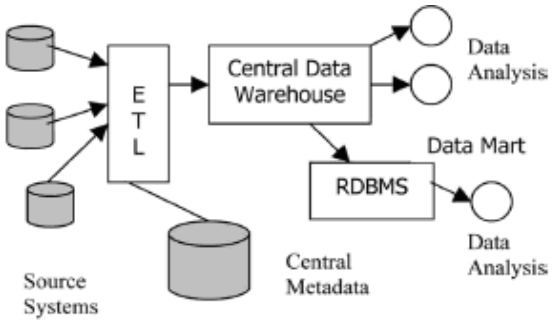

Figure 1. Central DW with Dependent Data Mart

Some of the best practices for DWA recommended by [7] are as follow;

1. Use a data model that is optimized for information retrieval.
2. Carefully design the data acquisition and cleansing processes for your DW.
3. Design a metadata architecture that allows sharing of metadata between components of your DW.
4. Take an approach that consolidates data into 'a single version of the truth'.
5. Consider implementing an ODS (Operational Data Store) only when information retrieval requirements are near the bottom of the data abstraction pyramid and/or when there are multiple operational sources that need to be accessed.
6. Create a capacity plan for your BI application & monitor it carefully.

3.2 Data Warehousing Process

The DWP tasks identified by [5] are as follow;

3.2.1 business requirements collection and analysis: It including interviews, joint application development (JAD), use of standard templates, requirements prioritization, use of subject areas, and review of existing documents.

3.2.2 data modeling: There are three levels of data modeling: conceptual design, logical design, and physical design.

3.2.3 data mapping: This process deals with the mapping of different data descriptions in the warehouse.

3.2.4 etl design: The set of functions that extract, transform, and load the data from the operational systems into the warehouse is called ETL.

3.2.5 end-user application design: Data warehouses are dedicated for providing information to business managers and executives. The information that they need is delivered through different types of end-user applications, including ad hoc queries, reports, OLAP functions, and data mining programs.

3.2.6 project planning: This process involves assessing the organization's DW readiness, scoping out the project, creating a project plan, identifying the resources available for the project, meeting critical deadlines, keeping costs within budget, supporting important functionalities, etc.

3.2.7 implementation: There are different approaches to implementing a DW in a target environment. They include the classical systems development lifecycle (SDLC) or "waterfall" approach, the iterative or "spiral" approach [1].

3.2.8 data quality management: It is the management function that determines and implements the data quality policy. A data quality system encompasses the organizational structure, responsibilities, procedures, processes and resources for implementing data quality management [9].

3.2.9 business continuity management: Most of the data warehouses are designed for supporting strategic decision-making. It is therefore quite important to provide plans to create archives and to develop recovery techniques in case there is a failure in the DW.

3.2.10 change management: Most data warehouses go through a lot of changes once they are installed. Changes in any DW can be hard to control. Data related to the maintenance of the warehouse are often dispersed and stored in a variety of formats. Change management is further complicated when the format of the source data are pushed onto the warehouse, rather than being controlled by warehouse administrators.

3.3 Steps-Toward Data Warehouse Architecture

Step 1: Always select the data model that is optimized for information retrieval, so that ETL(Extract,Transform,Load) process should be performed efficiently and accurately. The process for this step includes the requirement collection and analysis(i-e JAD, review of existing documents) and data modeling.

Step 2: Design the data acquisition and cleansing processes for your data ware house. Then the data coming from source systems will be clean and reliable to perform any action on it. The process are data mapping, that deals with the mapping of data of different descriptions

Step 3: Design a metadata architecture that gives the concept of sharing of metadata between components of your data ware house. The concept of meta data sharing here resolve or efficiently handle in this step is by meta data accessibility and knowledge timelines of dat. Meta data reporting enables to avoid the mistakes related to schema information.

Step 4: Operational Data Store implementation will be only when information retrieval requirements are at the lower level of the data abstraction pyramid and/or when there are multiple operational sources that need to be accessed. There are different ways to implement DW in target environment.

Step 5: There should be a data ware house monitor that evaluate the results of the queries. This process also enables the quality aware data warehouse architecture concept. It includes the system availability which is the percentage of time the source or data warehouse system is available.

Step 6: Integration of change is also a major issue during the maintenance of DW.

This process includes the following:

- Re-scoping DW development.
- Planning priorities.
- Redefining business objectives.

4. Comparison of Methodologies

The authors have analyzed 5 different DW methodologies, which they believe are fairly representative of the range of available methodologies. The sources of those methodologies can be classified into two broad categories: core-technology vendors (IBM, Microsoft, and Teradata) and information modeling companies (Creative Data). DW methodologies are rapidly evolving but vary widely because the field of data warehousing is not much mature. None of the methodologies discussed in this paper has achieved the status of a widely recognized standard as yet because every vendor thinks in their own specific domain but somehow everyone's target is same.

The comparison is based on the following benchmarks, which are the significant sub tasks of the DW development process [5]. These are the standard practices used by these vendors for;

- Requirement analysis

- Ø System development
- Ø Data modeling
- Ø ETL
- Ø Application design

Table 1 and table 2 summaries the result of our comparison based on the above mentioned benchmarks.

	Standard Practices for Req. Analysis	Standard Practices for Sys. Development
Creative Data	• Interview • JAD	• Iterative • Interview • Central DW with DM
Kimball	• Interview • Subject area • prioritization	• Dimensional life cycle • Interview • Subject area • Prioritization • ER-dimensional • Data mart
IBM	• Interview • JAD	• Iterative • Interview • Central DW with DM
Microsoft	• Interview • JAD	• Iterative • Interview • Central DW with DM
Teradata	• Interview • JAD • Template • Subject area	• Iterative • Interview • Template • Subject area • ER-relational • Central DW with DM

Table 1

	Standard Practices for Data Modeling	Standard Practices for ETL	Standard Practices for Application Design[5]
Creative Data	• ER-dimensional	• Immediate & Deferred • Source file • Timestamp • Automated • Incremental • Full refresh	• Client server & web MOLAP • Data mining
Kimball	• ER-dimensional	• Deferred • Timestamp • Automated • Incremental	• Client server & web ROLAP • MOLAP & HOLAP
IBM	• ER-dimensional	• Immediate & Deferred • Source file • Timestamp • Automated • Incremental • Full refresh	• Client server & web ROLAP • MOLAP & HOLAP
Microsoft	• ER-dimensional	• Immediate • DB trigger • Incremental	• Client server & web ROLAP • MOLAP & HOLAP
Teradata	• ER-relational	• Immediate • Source file • Automated	• Client server & web ROLAP • Data mining

Table 2

The most noticeable thing in this comparison is that the requirement analysis involves interview and JAD almost in all of the five methodologies. This emphasizes the importance and similarity of stakeholder involvement. Another important practice is that the system development process is iterative, which enable us to modify the system development processes by extending the DW architecture.

5. Quality Aware Data Warehouse Architecture

The role of data warehousing is to behave as a means of centralized information flow control, which is neglected by some of the organizations. As a consequence, a large number of quality aspects relevant for data warehousing cannot be expressed within the current data warehouse architecture.

In this section, authors discuss how to extend the architectural model of DW to support explicit quality models. Quality is a subjective phenomenon so authors must first organize quality goals according to the stakeholder groups that pursue these goals. On the other hand, "quality goals are highly diverse in nature. They can be neither assessed nor achieved directly but require complex measurement, prediction, and design techniques, often in the form of an interactive process" [6].

There exist different roles of users in a data warehouse environment. The Decision Maker usually employing for example an OLAP query tool to get answers interesting to him. User is mainly concern with the timeliness, ease of querying and the correctness of the result. The Data Warehouse Administrator needs facilities like error reporting, Metadata accessibility and knowledge of the timeliness of the data, in order to detect changes and reasons for them, or problems in the stored information. The Programmers of Data Warehouse Components can make good use of software implementation standards in order to evaluate their work. Metadata reporting can also facilitate their job since they can avoid mistakes related to schema information. Therefore, authors summarized the

quality dimensions of three stakeholders, the decision maker, data warehouse administrator and the programmer. [6]

5.1 Quality Goal

DW quality provides assistance to DW designers by linking the main components of DW reference architecture to a formal model of data quality [9].
The goal of our proposed architecture is first to understand, then controlling and improving the *Quality* of DW. Design and Administration Quality, Software Implementation Quality, Data Usage Quality and the most important Data Quality are the major dimensions of quality, discussed in [6]. Following components and roles are being added to develop a "Quality Aware Data Warehouse" in Figure 2.

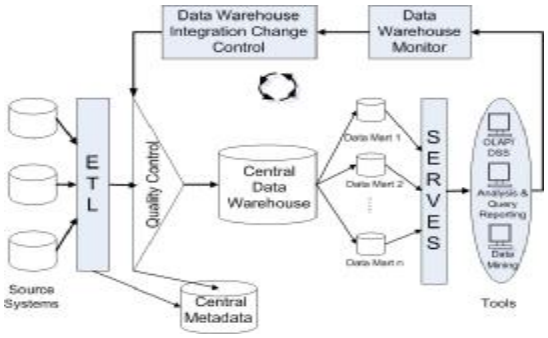

Figure 2. Quality Aware Data Warehouse Architecture

5.3 Quality Control Component

The approach begins by assuming the above mentioned dimensions are being fulfilled. Authors introduced a component named as "Quality Control" just after the ETL in the traditional DWA. This component is responsible for evaluating the quality of the data after performing ETL as well as analyzing that the data within the DW is model to represent adequately and efficiently the information available to it. This component will also justify the reason for aggregation, the freshness, completeness, accuracy, consistency and credibility of data. Another important consideration is that to avoid unnecessary redundancy during the source integration process.
Metadata management is a big challenge to many DW projects, mainly because there is much heterogeneity among tools and products for creating and managing metadata in a DW environment [3]. As Meta data is of high significance so after building it for the first time, it will be evolved by the quality control component. This evolution of Metadata is concerned with the way the schema evolves during the DW operation.

5.2 DW Monitor

Data Warehouse Monitor is responsible for evaluating the result of the queries. This component communicates directly with the decision maker using any of the tools like the DSS, OLAP or GIS etc.

It includes the following;

- The accessibility which is related to the possibility of accessing the data for querying.
- The security which describes the authorization policy and the privileges each user has for the querying of the data.
- The system availability which is the percentage of time the source or data warehouse system is available.
- The usefulness which describes the timeliness as well as the responsiveness of the system.

A DW by itself does not create value; value comes from the use of the data in the warehouse [10].

5.4 DW Integration Change Control

Integration change control is an important issue to consider while maintaining a DW. Surprisingly, very few vendors incorporate change management in their methodologies.
Authors have introduced a DW integration change control component because there are various changes that affect the DW. As our suggested architecture is not defined once in the beginning but emphasis on evolution, that is why there is a need of a separate team performing the activities related to it under their supervision.
Major consideration of this component involves re-scoping DW development, planning priorities, redefining business objectives and other related activities. Newer technologies could also affect the way an e-commerce site is set up and introduce changes [4].
Integrated change control involves identifying, evaluating, and managing changes throughout the life of DW. It is an important issue to consider while maintaining a DW. The main objectives includes, first evaluate either the change is required or not then perform the required actions to make that change within the DW repository and at last check the performance improvement that resulted after that change.

Introducing these new components within the existing architecture is a challenging task which is accomplished by hiring a separate team for maintaining the quality. Quality is not a one time process but it continues throughout the life of any project. In DW quality is very critical because decision making evaluates the future of any business. The team will be headed by a quality manager and other team members will be a part of the team. To

avoid conflicts and smooth running of this component proper and clear description of roles and responsibilities of the team members should be defined in advance.

6. Conclusions

In this study, authors reviewed the five current DWP standard practices followed in the industry and perform a comparison on the bases of some pre-defined benchmarks. Our study contributes an effort towards developing a centralized data warehousing process by highlighting the similarity of the existing methodologies.

Developing a DW is a software project, which is being made by following the traditional software project management practices. Software quality management is one of the sub-processes of SPM, where authors consider quality from the beginning of the project and it continues throughout the project life cycle. When authors made the architecture plan, quality plan is suggested to build at the same time. But the stakeholders are allowed to re-define their quality goals at any time. DW integration management then evaluates those suggestions and then performs the actual change within the system.

Further extension and refinement to this architecture is possible by implementing this model and evaluating the difference it makes. Based on this model as well as by combining the contributions of other researchers some standards can be defined that every DW needs to fulfill.

7. References

[1] W. H. Inmon, *Building the Data Warehouse*, 3rd ed. New York: Wiley, 2002.

[2] R. Kimball, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouse*, 2nd ed. New York: Wiley, 1996.

[3] L. Carneiro and A. Brayner, "X-META: A Methodology for Data Warehouse Design with Metadata Management," *in Proc. 4th Int. Workshop Design Manage. Data Warehouses (DMDW'02)*, Toronto, Canada, 2002, pp. 13–22.

[4] A. Sen and A. P. Sinha, "A Comparison of Data Warehousing Methodologies," *Commun. ACM*, vol. 48, no. 3, pp. 79–84, Mar. 2007.

[5] A. Sen and A. P. Sinha, "Toward Developing Data Warehousing Process Standards: An Ontology-Based Review of Existing Methodologies". *ieee transactions on systems, man, and cybernetics applications and reviews*, vol. 37, no. 1, january 2007.

[6] M.Jarke, M. A. Jeusfeld, C. Quix AND P. Vassiliadis, "Architecture And Quality In Data Warehouses: An Extended Repository Approach," *Information S.wtems* Vol. 24. No. 3, pp. 229-253. 1999.

[7] DW Architecture Best Practices, Cohesion Institute. www.irmac.ca/0506/DW Architecture Best Practices.ppt. 2008 May 2.

[8] S.Luj´an-Mora and J. Trujillo,"A Comprehensive Method for Data Warehouse Design," in *Proc. 5th Int. Workshop Design Manage. Data Warehouses (DMDW'03)*, Berlin, Germany, 2003, pp. 1.1–1.14.

[9] M. Jarke, Y. Vassiliou, "Data Warehouse Quality: A Review of the DWQ Project" *in Proc. 2nd Conference on Information Quality*. Massachusetts Institute of Technology, Cambridge, 1997.

[10] B. List, R. M. Bruckner, K. Machaczek, J. Schiefer, "A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse," *Springer-Verlag Berlin Heidelberg*, 2002, pp. 203–215.