International Journal of Grid Distribution Computing
Vol.8, No.2 (2015), pp.37-48
http://dx.doi.org/10.14257/ijgdc.2015.8.2.05

Challenges in Efficient Data Warehousing

Muhammad Arif[1,2] and Faheem Zaffar[2]

[1]*Faculty of Computer Science and Information Technology, University of Malaya 50603 Kuala Lumpur, Malaysia*
[2]*Computer Science Department, Comsats Institute of Information and Technology Islamabad Pakistan*

Abstract

Data warehousing has proved its importance as a source of decision support in competing market strategies and thus a lot of work is carried out in various aspects of it like conceptual enterprise perspective to ensure cost benefit is achieved by gaining competitive advantage which requires physical organization of data to improve the quality of data and quality of service. This paper gives insight how this is practically achieved, and where researchers are confronted with several challenges like privacy of data, access performance, efficient storage and cost benefit achievement.

Keywords*: Conceptual Enterprise Perspective, Physical Organization of Data, Dimensional-Modeling, Meta-Data Management, Quality of Data and Service*

1. Introduction

Data warehousing (DWH) has become the necessity for every business today. Now traditional RDBMS are not sufficient alone as market competition has increased and data warehousing as a source of decision support has proved its importance. DWH has its application in various areas like fraud detection, profitability analysis, direct mail/database marketing, credit risk prediction, customer retention modeling, yield management, inventory management. DWH is all about architectures, algorithms and tools used to extract data from heterogeneous sources (www data, semi-structured sources, archived data and operational databases) and transforming it to standard format and integrating it in a single repository which provides ad-hoc access to knowledge workers, thus helps in making effective decisions. For more efficient use of data warehouse, Online Analytical Processing (OLAP) tools are used to access data from DWH for complex data analysis (multi-dimensional data analysis) and decision support activities. DWH has evolved a lot in different aspects but still have issues like data quality and quality of service and cost benefit of data warehouse where researchers are putting their efforts. In order to understand the evolution and challenges it is necessary to understand the basic architecture of data warehouse.

ISSN: 2005-4262 IJGDC

Traditional Data Warehouse Architecture

Keeping in view traditional architecture, data warehouse is comprising source databases, materialized views, data transport agents those are responsible for shipping the data from one database to another and repository which is holding meta data to keep track of whole system and modifications done with passage of time. Since data comes from variety of sources in enterprise which are different in nature and format, and therefore a mechanism is required to transform that data in some uniform standard format. For this purpose of data extraction from heterogeneous sources component named ***wrapper*** plays its role in architecture of data warehouse. Another component know as ***mediator*** performs the responsibility of transforming data to resolve conflicts and integrating it in a single repository in the form of materialized views. Since the purpose of DWH is to provide ad-hoc access to knowledge workers, these materialized views are aggregated so that different groups of analyst may access them.

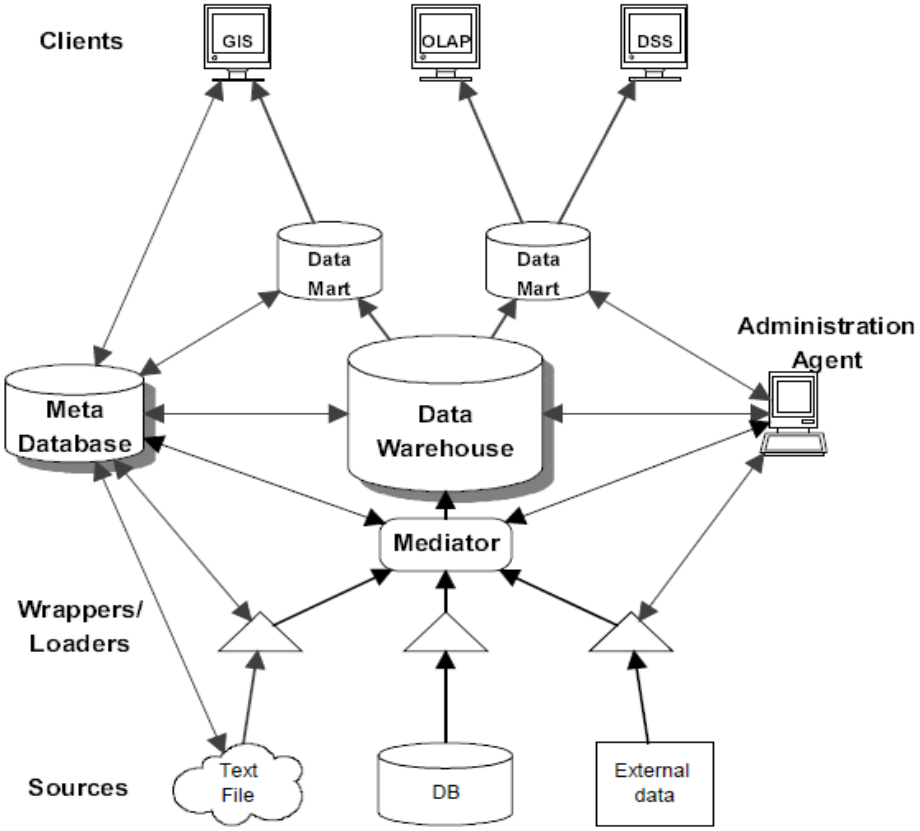

Fig. 1: Traditional Data Warehouse Architecture

Figure 1. Traditional Data Warehouse Architecture

There is concepts of ***data marts*** which are developed on the data in data warehouse with aggregates for specific domain of interest where they serve as second level of cache and decision support is provided by the use of special tools for data mining. A meta data is maintained about data warehouse which describes the use of data warehouse and its evolution [1].

2. Summaries

Conceptual Enterprise Perspective

Even after a lot of progress in evolution of data warehousing, still there are issues which need to addressed to make efficient use of data warehousing especially for cost benefit against managing much large data. The real challenge is that, is data warehousing keeping activity is aligned to goals of enterprise. In other words, the traditional architecture of data warehouse does not covers the aspects which may play major role in designing of data warehouse and thus may deliver actually desired benefit to enterprises. There is no picture of quality problem and management strategies covered in above architecture. Here comes, what is enterprise perspective to keep the data warehouse and how this perspective can be represented in above architecture[1]. In order to make better understanding, consider below figure.

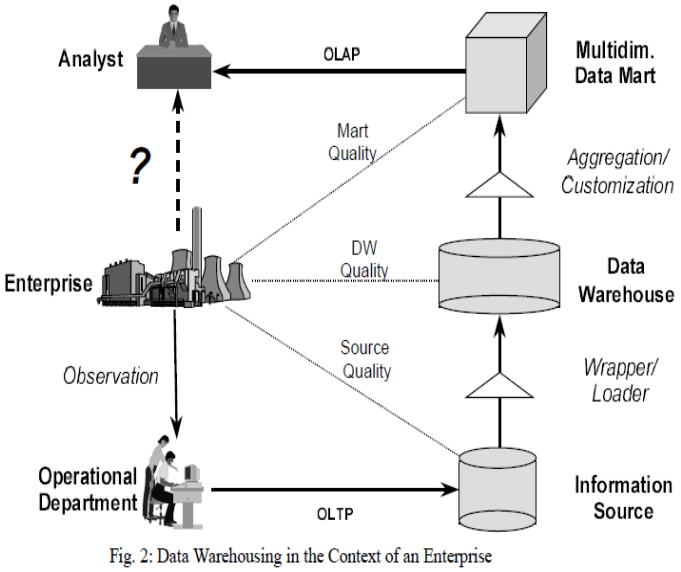

Fig. 2: Data Warehousing in the Context of an Enterprise

Figure 2. Data Warehousing in the Context of an Enterprise

The data warehousing related information gathering and maintenance explained in Fig 1 is available on right hand side while the process of using this information is mentioned on left hand side. Left hand side gives the picture about, how enterprise is interacting with data warehouse. Let's go in little more depth, if an enterprise is spending 30% of the revenue on data warehouse keeping and the actual benefit gained with its usage is just 5% then it is alarming situation and analyst must be looking for the clue that what is wrong going on with the business. As he cant view the business directly and totally relying on the reports provided by the operational departments but that might not be sufficient to know what is wrong and where is wrong. Question can still be raised that organization has spent a lot on data warehouse but why still results are not very impressive. One possible answer is that there is no or very little compatibility of data managed with enterprise view of operational departments. Reason can be wrong aggregation in the materialized views and there can be many other reasons. Solution of this problem is that there must be an model that caters the conceptual relationships between enterprise model, departmental OLTP sources and OLAP clients used for decision making [1].

The main point of above discussion is that data warehousing is supposed to be the mean of centralized information flow control. But in traditional architecture, a large number of quality aspects relevant for data warehousing cannot be expressed with the current DW Meta models. This problem is tackled by handling it in two steps. First step is

make efficient use of Meta data maintained about the data warehouse architecture through use of enterprise model. "Second step involves the use of different mathematical techniques for measuring or optimizing certain aspects of DW quality are being developed" [1].

Data warehouse has metadata repositories which contain information track of activities performed at back-stage, concerned with refreshing of data with clean, up-to-date and structurally reconciled data, also information about the contents, their structure and location, information on the infrastructure and physical characteristics of components and the sources of the data warehouse, and, information including security, authentication, and usage statistics that aids the administrator tune the operation of the data warehouse as appropriate. There is no single decidable formalism that could cover the handling of all these aspects uniformly in a meta database. [1]

To practically achieve this, architecture is divided into three perspectives

A. Conceptual Perspective

The conceptual perspective serves as a source of information system from enterprise perspective. Enterprise model integrates the all related conceptual objects whereas client model and system of source information are presented as views on the enterprise model. The main goal is to achieve independence from physical organization of data so that relationship between inter-related concepts can be analyzed by intelligent tools. Thus integration of multiple information sources can be done in smooth way [1].

B. Logical Perspective

Logical perspective focus on actual data models involved in data warehousing. The basic element of logical perspective is Schema and it is related to physical component which implements the logical schema. This extended architecture supports multiple data models like relational model, multi-dimensional model and object oriented data models.

C. Physical Perspective

Physical perspective refers to the software and other physical components that are required for establishing data warehouse.

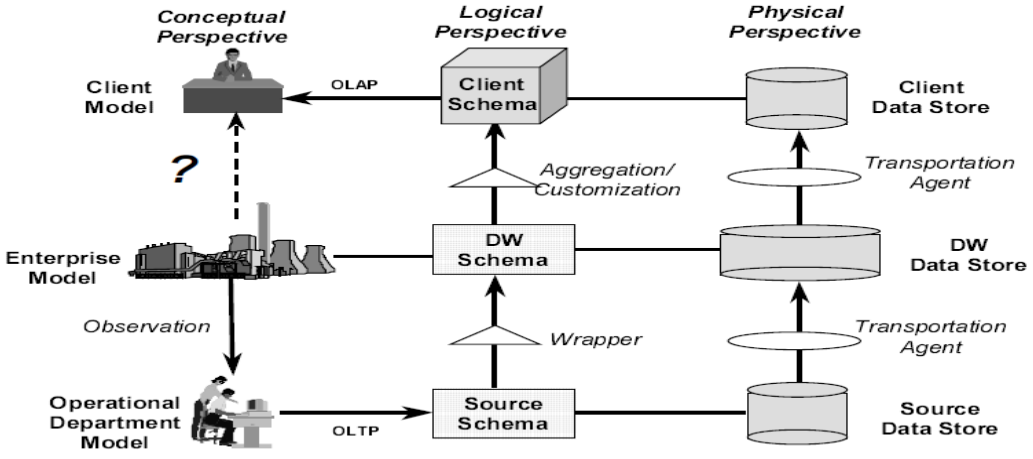

Fig. 3: The Proposed Data Warehouse Meta Data Framework

Figure 3. The Proposed Data Warehouse Meta Data Framework

Meta Data Standards

To better meet the complexity of data structures and processes deriving the data warehousing, a consistent management of metadata is required. Metadata is classified into business meta data which is used by end-users and technical metadata produced and used by administrators or by other software component of data warehouse[2].

Open Information Model (OIM) and the **Common Warehouse Meta model** (CWM) specification are two accepted standards for metadata representation and exchange, proposed by Meta Data Coalition and OMG. To clarify a standard, for metadata representation, it is required that the complete description of a Meta model is available with all its elements, their semantic contents and interdependencies between elements. The standard is inherently independent of any specific implementation. A standard for exchange is based on a unique Meta model as well but it also contains the definition of ready-to-use interfaces that specify the meta model in a wide-spread language like XML or CORBA IDL [2].

Open Information Model → The purpose of OIM is to support tool interoperability across technologies and companies via a shared information model. OIM is designed to encompass all phases of information systems development, from analysis through deployment. OIM version 1.0 was adopted in July 1999, by the Meta Data Coalition as a standard. The Meta Data Coalition aims at the definition, implementation and ongoing evolution of a metadata interchange format standard and its support mechanisms [2].

Common Warehouse Meta model →the main purpose of **CWM** is to enable easy interchange of common warehouse metadata between warehouse tools and warehouse metadata repositories in distributed heterogeneous environments and thus considerable documents are provided with the CWM Meta model expressed in XML and interfaces generated as CORBA IDL. CWM is based on the following OMG standards: UML, MOF (Meta Object Facility), a Meta modeling and meta-data repository standard, and XMI, an XML-based standard for metadata exchange originating from OMG [2].

Comparison A coarse comparison of the two competing specifications

OIM	CWM
OIM is designed to encompass all phases of information systems development and contains one package that is specifically intended for data warehousing, the Database and Warehousing Model.	CWM deals entirely with metadata for data warehousing only and provides a framework for representing and exchanging metadata about data sources, data targets and warehouse processes that create and manage them.
Result → OIM is much broader in scope than CWM.	
OIM uses UML 1.3 as their foundation but OIM is not compliant with the Meta Object Facility (MOF) standard.	CWM uses UML 1.3 as their foundation but CWM is compliant with the Meta Object Facility (MOF) standard.
Result → CWM has an inherent repository orientation provided by the MOF compliance (e.g. representation of the meta data as CORBA objects). OIM, on the other hand, is not a specification of a repository API or implementation; it focuses on the description of information, not on data access and management.	

The differences and similarities between OIM and CWM can be summarized as follows:

1. They build on the same ground, namely UML and XML, raising the hope that a unification of the two models may someday be possible.
2. They differ in scope: while CWM is primarily restricted to data warehouse metadata, OIM has a broader purpose.

3. They do not provide mechanisms to categorize information according to different conceptual levels of metadata, but both standards provide a package structure to separate different warehouse areas.
4. They provide support to represent and exchange core warehouse metadata with an emphasis on the technical metadata. For the time being, both standards fall short in providing mechanisms for handling business metadata.
5. In contrast to CWM, OIM is currently oriented and limited to the support of relational data (*e.g.*, in transformations, OLAP). While CWM provides a sophisticated mechanism to exchange metadata between tools or repositories relying on the XMI standard, the proposed solution for OIM is rather hand-knitted.
6. Despite the fact that the OMG and the MDC established a formal technical liaison (the MDC is now a Platform Member of the OMG, and the OMG is a member of the MDC). In order to build consensus on metadata standards", significant efforts are required to unify OIM and CWM.

It is likely that the two standards will co-exist for some time and metadata exchange between the two standards (via XML) will be commonplace. However, to support a complete integration of all warehouse metadata stored in repositories compliant with OIM or CWM, a unified and extended model will be absolutely necessary[2].

Physical Organization Of Data

Physical organization is a major role playing aspect in terms of efficient storage and access performance. In this context various techniques like transposition, compressing multidimensional space, view materialization methods, pre-partitioning the data cubes, data cube update techniques are used [6].

Transposition

Transposition is sometimes referred to as a “vertical partitioning” of a table. Clearly, the main advantage of this approach is that for summary operations and statistical query (such as cross tabulations) only the relevant columns need to be retrieved. This improved greatly the access efficiency of statistical queries [6].

Multidimensional Space

The most obvious idea for compressing multidimensional space is “array linearization”. This term refers a fairly well-known calculation of the position of a cell in a multidimensional array, based on number of possible values in each dimension. Thus, instead of storing the columns of the tabular representation we need to store the distinct values of dimensions only once. This is the main idea behind Multi-dimensional OLAP products. Using array linearization works well when the multidimensional space is dense, i.e. that each cell has a measured value associated with it. However, in many cases many of the cells have nulls or zeros [7].

View materialization methods

For very large datasets (in the gigabyte range) it is expensive to calculate summarizations, since too much data has to be retrieved from secondary storage. Since SDBs and OLAP databases are often stable, *i.e.*, have very few updates if any, pre-generating and storing summarization views is a worthwhile possibility to consider. This is a problem of space-time tradeoff. Given a limited extra space, which summarizations should be pre-calculated for the maximum access benefit [5].

Pre-partitioning the data cube

Another idea for improving access efficiency is to take advantage of "range" type queries. Realizing that many of the queries involve "slicing" and "dicing", it makes sense to pre-partition the data cube into sub-cubes. The advantage of this approach is that only the sub-cubes that contain data relevant to the query or summarization request need to be read from secondary storage. For a given query that spans multiple sub-cubes, each sub-cube may contain part of the relevant data. Consequently, the access software needs to manage multiple sub-cubes and assemble the required data from them. If there is no information about the access patterns to the data cube, a symmetric partitioning is appropriate; that is, all the sub-cubes are of equal sub-dimensions. The only parameter to be determined is the size of a sub-cube. This can be determined by estimating the average size of a query[5].

Data Privacy

Data privacy is major concern of every organization and it can be really challenging to achieve. This section highlights various aspects of data privacy and their solutions in context of data warehouse. It is usually permissible to make public summarized data, but not the individual data used for generating the summary. Census data is an obvious example, but other domains, such as criminals' records, or financial data about individual companies, need to be protected by agreement or by law. The problem is whether privacy protection can be guaranteed. This problem is referred to as the 'statistical inference" problem [6].

There was a lot of work done in this area in the 1980; many references can be found in. However, it is worth mentioning here an important negative result. In this paper it was shown that it is always possible to compromise a database by using a combination of queries. In fact, the authors have developed a procedure (called a "tracker") for finding the collection of queries to infer a desired fact. One possible solution seems to restrict the number of individuals included in the response to a statistical summary query, called a "query set". For example, we may choose to restrict the response to "average salary", only if the average is for 5 people or more. We say that the query set is restricted to be greater than 5. Restricting query set size if not enough, since intersecting query sets can compromise information. Thus, it is necessary to monitor the size of query set intersections as well. Consequently, many other approaches were proposed.

(i) Limiting the query set intersection. This requires keeping track of all query sets, and making sure that a new query set does not intersect with previous one to produce a query set size below the permitted size. This can be used for small databases, but eventually it reaches the point that no new queries can be answered.

(ii) Random sample from a query set. This is useful for very large datasets, when the typical query set is large.

(iii) Pre-partition the dataset into cells, and give responses that involve whole cell only. This approach is used fairly effectively with the Census data in the US, but requires "cell suppression" (*i.e.*, some cells that contain too few individuals are cannot be reported).

(iv) Input data perturbation -- stored statistically correct, but perturbed data for general consumption.

(v) Output data perturbation -- perturb results given to users within some statistical limits.

From above scenarios it is clear that there are no easy solutions to the privacy problem. All the solutions proposed have some disadvantages. However, given the importance of privacy, an imperfect solution is better than none. It is interesting to note that although

privacy needs to be supported in many OLAP applications, they are currently ignored in OLAP systems and literature. The interest in privacy continues in the SDB community, such as the Conference on Data Protection [6].

Quality of Data and Service [4]

Quality of data and service is the basic necessity of data warehouse as it serve as the decision making source. To gain competitive advantage decision making process is completely relying on the quality of data and services. The main difference in this approach resides in the following points:

(i) A clear distinction between subjective quality goals requested by stakeholder and objective quality factors attached to DW objects.
(ii) Quality goal resolution is based on evaluation of the composing quality factors, each corresponding to a given quality question.
(iii) Quality questions are implemented and executed as quality queries on the semantically rich metadata repository.

The refreshment process is one of the main DWs processes for which the quality is an important issue. The associated quality template includes quality dimensions such as coherence, completeness and freshness.

- *Data coherence* → the respect of (explicit or implicit) integrity constraints from the data. For example, the conversion of values to the same measurement unit allows also doing coherent computations.
- *Data completeness* → the percentage of data found in a data store, with respect to the necessary amount of data that should rely there.
- *Data freshness* → the age of data (with respect to the real world values, or the date when the data entry was performed).

Table 1

Paper no	Author name	Published date	conference / journal paper	DWH Type	Tool name	Challenge Handled
[1]	MATTHIAS JARKE, MANFRED A. JEUSFELD, CHRISTOPH QUIX, PANOS VASSILIADIS	1998	Journel (INFORMATION SYSTEMS -OXFORD-PERGAMON PRESS)	Traditional	ConceptBase	Quality of Data
[2]	Vetterli, A Vaduva, M Staudt	2000	Conference (ACM Sigmod)	Centrelized	Tool Independent	Meta Data Management
[3]	CHARLES BONTEMPO, GEORGE ZAGELOW	1998	Journel (COMMUNICATIONS OF THE ACM)	Centrelized / Distributed	IBM DB2	Quality of Data and Cost Benefit
[4]	Panos Vassiliadis, Mokrane Bouzeghoub, ChristophQuix	1998	Journel (INFORMATION SYSTEMS -OXFORD-PERGAMON PRESS-)	Centrelized	ConceptBase	Quality of Data
[5]	Ralph Kimball, Margy Ross	2006	Wiley Computer Publishing		No tool discussed	Dimensional Modeling
[6]	Arie Shoshani	1997	Conference (ACM Sigmod)	Centrelized	No tool discussed	Efficient storage and access performance and data privacy
[7]	Harinarayan V., Rajaraman A., Ullman	1996	Conference (ACM Sigmod)	Traditional	Lattice Framework	Implementing Data Cube Efficiently

Table 2

Paper no	Experimental environment	Related architecture / work	Future idea
[1]	Traditional DWH	Previous projects were focused on source integration aspects like databases, SGML documents, or unstructured files	Systematic quality analysis and quality-driven design
[2]	Traditional/Centralized/Distributed DWH	Object-Oriented framework for distributed applications	Consistent meta management for achieving data quality
[3]	Multi-database servers	scalability, performance, wider windows of system availability, and a strong skills base	cross-functional information for achieving enterprise level information requirements
[4]	Traditional DWH with Extended Repository Model	Meta Quality Model and Goal-Question-Metric (GQM) to achieve quality of data and service	Improving in proposed Meta Quality Model
[5]	MOLAP Implementation	Dimensional Modeling Aspects	Efficiency of dimensional modeling techniques
[6]	Traditional 2-dimensional (2-D) representation of SDBs, and data cube representation of OLAP databases	Multidimensional datasets with statistical summarizations over the dimensions of the data sets	Conceptual modeling of the data and operations over them, efficient physical organization and access methods, as well as privacy issues.
[7]	Multi-dimensional databases (MDDB)	Improvement in managing multi-dimensional data	Implementation details for making Data Cube

Table 3

Paper no	Gui Schema Yes/no	Data warehouse approach	Data ware house type	Application type	Analysis type	Layered approach	Tool	ETL Yes/no
[1]	Telos	OLAP	Generic	Generic	Data Quality	5 layers	Goal-Question Matrix	Yes
[2]	N/A	Generic	Generic	Generic	Generic	N/A	N/A	N/A
[3]	N/A	Generic	Generic	Generic	...	6 layers	N/A	Yes
[4]	N/A	OLAP	Generic	Generic	Data Quality	3 layers	*Statistical Process Control*	Yes
[5]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
[6]	N/A	OLAP	Financial	N/A	N/A	N/A	N/A	No
[7]	Lattice Diagram	OLAP	Generic	Generic	Multi-Dimensional Data	N/A	Lattice Framework	No

Table 4

Paper no	OLAP technique name	DWH scope	Noval idea or modification	No of star schemas	Technique name	Evaluation	Validation	SQL/ ORACLE	Motivation
[1]	MOLAP	Enterprise	Meta Repository	No	Goal-Question Matrix	Meta Quality	Quality Factor	ConceptBase	Conceptual, Physical and Logical Perspective of Enterprise
[2]	N/A	Generic	N/A	N/A	Common Warehouse Model and Open Information Model	N/A	N/A	CWM -> SQL1999 OIM -> SQL-92	Comparison of Techniques
[3]	MOLAP / ROLAP	Enterprise	Generic DWH Architecture	N/A	N/A	N/A	N/A	DB2, Essbase	Cost Benefit of DWH
[4]	MOLAP	Enterprise	Meta Repository	N/A	Goal-Question Matrix	Total Quality Management Approach	Goal Question Matrix	ConceptBase	Conceptual, Physical and Logical Perspective of Enterprise
[5]	MOLAP	Generic	N/A	N/A	N/A	N/A	N/A	N/A	Book on DM
[6]	MOLAP	Generic	N/A	N/A	N/A	N/A	N/A	N/A	Comparison of DHW and Statistical Databases
[7]	MOLAP	Generic	Data Cube Efficiency	No	Hypercube lattice	TPC-D Bench Mark	TPC-D Bench Mark	MDDB	Implementing Efficient Data Cube

3. Conclusion

In order to get maximum money saving benefit from data warehouse, necessary things are that, don't get stuck in proprietary solutions, reduce complexity by following standards, identify the availability requirements those will be necessary and data warehouse must meet in few years[3]. Real strength of data warehouse can be exploited if quality of data and quality of service but are available in it otherwise it will not be helpful in decision making and instead of giving cost benefit it will result in loss. To achieve data quality meta data about the DWH plays important role and that can be improved by considering conceptual, physical and logical perspective of enterprise.

4. Future Work

Meta model described for quality of data and quality of service is partially validated. Software tools and applications are required; those can fully validate the proposed model.

References

[1] M. Jarke, M. A. Jeusfeld, C. Quix and P. Vassiliadis, “Architecture and quality in data warehouses”, Advanced Information Systems Engineering, Springer Berlin Heidelberg, **(1998)** January, pp. 93-113.
[2] T. Vetterli, A. Vaduva and M. Staudt, “Metadata standards for data warehousing: open information model vs. common warehouse metadata”, ACM Sigmod Record, vol. 29, no. 3, **(2000)**, pp. 68-75.
[3] C. Bontempo and G. Zagelow, “The IBM data warehouse architecture”, Communications of the ACM, vol. 41, no. 9, **(1998)**, pp. 38-48.
[4] P. Vassiliadis, M. Bouzeghoub and C. Quix, “Towards quality-oriented data warehouse usage and evolution”, Information Systems, vol. 25, no. 2, **(2000)**, pp. 89-115.
[5] R. Kimball and M. Ross, “The data warehouse toolkit: the complete guide to dimensional modelling”, Nachdr.]. New York [ua]: Wiley, **(2002)**.

[6] A. Shoshani, "OLAP and statistical databases: Similarities and differences", Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, ACM, **(1997)** May, pp. 185-196.
[7] V. Harinarayan, A. Rajaraman and J. D. Ullman, "Implementing data cubes efficiently", ACM SIGMOD Record, vol. 25, no. 2, **(1996)**, pp. 205-216.

Author

Muhammad Arif is a PhD student at Faculty of CS and IT, University of Malaya. Currently he is working on Medical image Processing. His research interests include image processing, E learning, Artificial intelligence and data mining. He joined UM as a Bright Spark Scholar in September 2013 for the period of 3 years. Before this he completed masters and bachelor degrees in Pakistan. He received his BS degree in Computer Science from University of Sargodha, Pakistan in 2011. He obtained his MS degree in Computer Science from COMSATS Islamabad 2013 Pakistan.