

# **Extension of GDI Approach to Ensure Quality for Data Warehouse Using Quality Manager**



*MS Research Dissertation by:*

**Arooba Hanif**

**(525-FBAS/MSCS/S09)**

*Supervised by:*

**Prof. Dr. Maqbool-Uddin Shaikh**

*Co-Supervised by:*

**Ms. Zareen Sharf**

Department of Computer Science  
Faculty of Basic and Applied Sciences  
International Islamic University Islamabad

**2011**



Accession No. TH-8584

MSC  
005.72  
ARE

- 1 - Data Record formats
- 2 - Data Conversion

DATA ENTERED  
*Aug 8 15/3/13*

**Department of Computer Science & Software Engineering**  
**Faculty of Basic and Applied Sciences**  
**International Islamic University Islamabad**

Date: 19-10-2011

**Final Approval**

This is to certify that we have read the thesis submitted by **Arooba Hanif, Registration No. 525-FBAS/MSCS/S09**. It is our judgment that this thesis is of sufficient standard to warrant its acceptance by International Islamic University, Islamabad for the degree of **MSCS**.

Committee:

**External Examiner:**

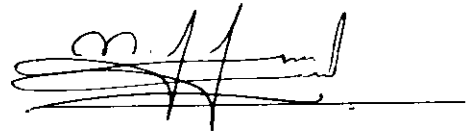
*Dr. Jamil Ahmed*  
*Professor/Dean*  
*Department of Computing & Technology,*  
*Iqra University, Sector H-9, Islamabad.*



---

**Internal Examiner:**

*Mr. Muhammad Imran Saeed*  
*Assistant Professor,*  
*DCS&SE, FBAS, IIUI.*



---

**Supervisor:**

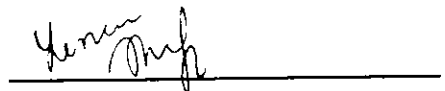
*Prof. Dr. Maqbool-ud-din Shaikh,*  
*Department of Computer Science*  
*Comsats, Chak Shahzad, Islamabad*



---

**Co-Supervisor:**

*Mrs. Zareen Sharf Khan,*  
*Assistant Professor*  
*DCS&SE, FBAS, IIUI*



---

# **Dedication**

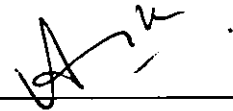
**Dedicated To My Beloved Parents Especially  
To My Mother**

**Arooba Hanif  
525-FBAS/MSCS/S09**

**A dissertation Submitted To  
Department of Computer Science,  
Faculty of Basic and Applied Sciences,  
International Islamic University, Islamabad  
As a Partial Fulfillment of the Requirement for the Award of the  
Degree of MS in Computer Science.**

## Declaration

We hereby declare that this Thesis "*Extension of Goal Decision Information Approach to Ensure Quality of Data Warehouse Using Quality Manager*" neither as a whole nor as a part has been copied out from any source. It is further declared that we have done this research with the accompanied report entirely on the basis of our personal efforts, under the proficient guidance of our teachers especially our supervisor *Prof. Dr. Maqbool-ud-din Shaikh*. If any part of the system is proved to be copied out from any source or found to be reproduction of any project from any of the training institute or educational institutions, we shall stand by the consequences.



---

Arooba Hanif

525-FBAS/MSCS/S09

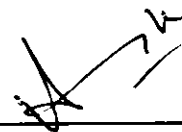
## Acknowledgement

First of all I am obliged to Allah Almighty the Merciful, the Beneficent and the source of all Knowledge, for granting us the courage and knowledge to complete this Project.

I owe my deepest gratitude to my **Parents**, my sisters **Lubna, Rabia** and brother **Adil** for their unflagging love and support throughout my carrier. Many thanks to my husband **Naeem**, who is a great hope to help me cope with any difficult situation, and my dearest **Grand Mother** who have supported me spiritually.

I would like to express my warm thanks to my Supervisor, **Dr. Maqbool Uddin Shaikh & Ms. Zareen Sharf**, without their guidance this research work would not have been presented.

There are number of people to whom I am greatly indebted without them this thesis might not have been completed, I want to thank to all my Teachers especially **Sir Qaiser, Sir Imran Saeed** and my friends especially **Atika Qazi**.



---

**Arooba Hanif**  
**525-FBAS/MSCS/S09**

## Project In Brief

<b>Project Title</b>	<b>Extension of GDI Approach to Ensure Quality for Data Warehouse Using Quality Manager</b>
<b>Undertaken By:</b>	<b>Arooba Hanif</b>
<b>Supervised By:</b>	<b>Prof. Dr. Maqbool-Ud-Din Shaikh</b>
<b>Start Date:</b>	<b>12-03-2010</b>
<b>Completion Date:</b>	<b>19-04-2011</b>
<b>Tools &amp; Technologies</b>	<b>Meta data mathematical calculations using Mathematical expressions</b>
<b>Documentation Tools</b>	<b>MS Office 2003/ 2007</b>
<b>Operating System:</b>	<b>Windows XP</b>
<b>System Used:</b>	<b>Pentium IV</b>



## Abstract

Data warehouse System as enterprise-wide is the collection of highly integrated data coming from various data sources (heterogeneous sources). Integration of data from heterogeneous sources, multiple programs and tools used for loading data into the Data warehouse makes the whole system very complex. And data warehouses primary purpose is of analysis therefore, it's very important to ensure its quality to make the quality analysis for its users.

In this view quality is a major factor for data warehouse not in terms of data only but also in terms of services provided by it. To handle the complexity of data warehouses due to its dynamic nature it greatly depends on meta-data. Meta data management support users to overcome such kind of complexities arise due to dynamic nature of data warehouse. This paper contributes new methods in the existing GDI (Goal Decision Information) model of Meta data management thus, solving the problems of available solutions.

# Table of Contents

## 1 Introduction

1.1	Introduction.....	1
1.2	Motivation.....	2
1.3	Background.....	4
1.4	Research Problem.....	5
1.5	Proposed Approach.....	5
1.6	Thesis Outline.....	7

## 2: Literature Survey

2.1	Introduction .....	9
2.2	Data Flow.....	9
2.3	Data Quality Review in Data warehouses.....	10
2.4	Data Quality Tools.....	11
2.4.1	Data Auditing Tools.....	11
2.4.2	Data Cleansing Tools.....	11
2.5	Data Migration Tools.....	13
2.6	DW Critical Success Factors.....	13
2.7	CSF in Data Warehouse Technology.....	13
2.8	Meta Data Repository.....	19
2.9	Related Work.....	22
2.9.1	Literature Survey.....	25

## 3: Requirement Analysis

3.1	Introduction .....	32
3.1.1	Meta data repository.....	33
3.1.2	Meta data types.....	33
3.2	Meta data & Data warehouse.....	34
3.3	Critical Scenarios.....	35
3.4	Focus of Research .....	36
3.5	Summary.....	37

## 4: System Design

4.1	Introduction .....	40
4.2	Design Requirements .....	41
4.3	Reference Architecture.....	41
4.4	The Proposed Architecture.....	44
4.5	Flow Chart Diagram.....	47
4.6	Algorithm.....	49
4.7	Summary.....	50

5: Implementation Support

5.1 Implementation support for extended quality model.....52  
5.2 Defining & Redefining goals.....52  
5.3 Quality Measurement.....54  
5.4 Case Study.....55  
5.5 Evaluation of the quality goal.....56  
5.6 Improvement Phase.....58  
5.7 Summary.....59

6: Conclusion & Future work

6.1 Comparisons..... 61  
6.2 Conclusion & Future Work.....64

## LIST OF FIGURES

Figure 1.1	<b>Traditional Data Warehouse Architecture [9].....</b>	<b>4</b>
Figure 2.1	<b>Data Flow [14].....</b>	<b>9</b>
Figure 2.2	<b>Data Flow &amp; Data Quality tools [14].....</b>	<b>10</b>
Figure 2.3	<b>Success Factors in FAC Corporation [1].....</b>	<b>14</b>
Figure 2.3.1	<b>Data warehouse CSF [1].....</b>	<b>15</b>
Figure 2.4	<b>Goal Question Matric (GQM) Approach [02].....</b>	<b>22</b>
Figure 2.5	<b>Extended Goal Question Metric Approach (GDI) [17].....</b>	<b>23</b>
Figure 2.6	<b>Goal Decision Information (GDI) Approach [9].....</b>	<b>24</b>
Figure 4.3	<b>The Goal-Decision Information Model [2].....</b>	<b>29</b>
Figure 4.3.1	<b>GDI Quality model [2].....</b>	<b>30</b>
Figure 4.4	<b>The Proposed Quality Model using GDI approach.....</b>	<b>32</b>
Figure 4.5	<b>Flow Chart.....</b>	<b>34</b>

**CHAPTER NO. 1**  
**INTRODUCTION**

## 1.1 Introduction:

Data warehouses are mainly used to retrieve the decisional sort of information and they are basically design to support the decision information system which usually based on integration of highly heterogeneous sources to extract, transform, aggregate data to facilitate the ado queries to extract the required information.

So we can say Information systems derived from heterogeneous databases are nowadays called as modern information system. When we integrate heterogeneous database to a single data warehouse then the most known issue confronted data warehouse system is 'low quality data'. Issues related to data quality problem can be identifying through data mining techniques like clustering, subspace clustering and classification of data.

Data from heterogeneous sources which are highly aggregated can be very complex warehouses which mainly deliver decisional support data to decision makers. Now most of the organizations are inclined towards Data Warehouses to retrieve information from enormous amount of raw data, so for the Top level managers data warehouse comes up like the key trend as they can have accurate and relevant information to support their decisions and improve strategic decisions.

For an effective Data warehouse system, the informational & technical aspects of quality of DW system should be properly incorporated at all the stages in development. As the quality & Correctness of data found among heterogeneous databases has never been an easy issue in current age where we have overwhelming amount of data.

## 1.2 Motivation:

With the advent of modern data analysis tools and the growing concepts of Business Intelligence, it has become radically important for top management to use historical data for quick analysis and balanced decision making.

For a business to collect and analyze data, and subsequent decision making a central repository for any business is need of the hour. Organizations either have converted their traditional storage to DW or in the process of doing that as data repository system for enterprises. It is best solution for decision making process in a business organization and business intelligence. Top level management of the organization requires historical data to keep track the business trends in the market.

Data warehousing is up to date architectures for strategic decision making. It gives tools for executives of business to methodically organize, understand, and use data to make advanced decisions. It has been analyzed using OLAP techniques that Data warehouse is an infrastructure used to extract, clean and store large amount of business data from different operational systems into a common storage format for productivity enhancement and efficiency.

Operational Data Store (ODS) consisted of a data store of volatile and integration of low granularity data. At ODS level, we usually perform transformation and cleaning processes so that clean and homogeneous data populated to Data warehouses.

In data warehouses we have local or client level views on top layer which has highly aggregated data, mostly derived from global warehouse. OLAP databases and data marts are called as kind of local warehouses which can be relational database systems or maybe multidimensional data structures.

So DW consisted of various components involving many stakeholders, with different goals and is monitored constantly by administrator using administrative tools. To administer the data ware components and processes, data should be tracked from a meta data repository as Meta Data repository basically acts as a way to keep the track or traces of all the records, design choices and keeps history of all the changes which are performed on its architecture and components.

Therefore, the good place to represent to represent quality goals of stakeholders explicitly is the Meta database of a data warehouse. Data warehouse quality can be analyzed through queries on the Meta database as Meta database holds the result of quality measurements

Thus, with the importance of data warehouse it is very essential to assure its quality as it acts as a tool for strategic decisions. However, focus has not given much on design and analysis of the quality of the data warehouse and it's a great problem from users Perspective. For such issues different techniques and model has been presented to improve the quality of data ware houses.



### 1.3 Background:

Traditional Architecture model of data warehouse constitutes various layers of data in which data from layer is derived from another layer. The lowest layer comprise of data that is derived from operational data sources. They may include structured or unstructured or may be semi-structured data stored in files [2]. The next layer of the architecture is primary data warehouse, also termed as global data warehouse. The global DW keeps a historical record of data that result from the transformation, integration, and aggregation of detailed data found in the data sources [2].

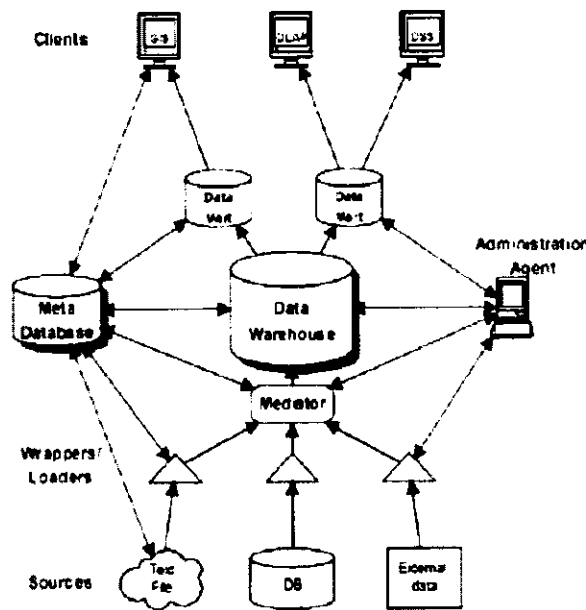


Figure 1.1: Traditional Data Warehouse Architecture [9]

So far different life cycles and techniques have been proposed for data warehouse development however development of data warehouse is still very complex and risky task. It is essential to store quality data in the data ware house as quality of the data warehouse became the main tool for strategic decisions [2]. However, the design and

analysis of the quality of a data warehouse is not well understood and is a great problem from the perspective of the users [2].

#### **1.4 Research Problem:**

To tackle the design and analysis of the quality of the data warehouse, a rich semantic data model was proposed [2] for the components of a data warehouse linked to a quality model using GQM approach. The quality framework adapts the Goal Decision Information GDI model to improve and quality assurance of the data warehouse model.

#### **1.5 Proposed Approach:**

In the proposed approach, emphasis will be given on quality goal, where a stakeholder has to manage the quality of the data warehouse, or a part of it. For example 'achieve the availability of source 's1' at least once per week in the viewpoint of DW administrator. The purpose of the Goal is obtained from the policy and the strategy of the organization. Quality criteria are used as abstractly different aspects of quality and are obtained from the policy and strategy of the organization. So this policy and quality criteria is used to define quality measurements and making use the information stored. This model assumes that the acceptable values, stored in the Meta data repository are provided by the stakeholder.

In the model focus would be on quality measurements in choosing the right source databases and the some quality queries which are executable queries are applied on the Meta database at any instance of time, because the Meta database is not just the CASE repository but an integral part of the runtime data warehouse system. The metadata repository acts as a path to trace all the design choices and a history of changes performed on its architecture and components.

So in the support of above matter emphasis is given to improve the quality of the data warehouse, in this regard so much work has been done and different models are presented to achieve the standard to improve the quality of the data ware houses.

Existing model that was presented in 2009 was extension of Goal decision quality model which achieved performance up to the some extend but common issue related to it is there is no suitable strategy given for materializing the quality measurements which is missing and also measuring the quality can be computationally expensive.

In the proposed approach a suitable strategy will be given to store quality data in the data ware house and to perform quality queries on the data repository. A certain criteria for quality measurements would be taken to avoid the cost of computations which could be expensive for materializing the quality measurements.

## **1.6 Thesis Outline:**

There are total six chapters in this thesis. In the first chapter detailed introduction about the data ware house and its quality is given and background of the problem. The second chapter is related to literature survey in which main emphasis is given on literature of quality models of data warehouses and the techniques and concepts previously used to enhance its quality.

In chapter three, focus is on requirement analysis to explore and explain the problem domain in detail, and what will be focus of this research. Chapter four is about system design, in which design requirement, architecture and methodology is discussed. And in the fifth chapter implementation of the research paradigms is given. In the sixth chapter conclusion & Future work is given.

**CHAPTER NO. 2**  
**LITERATURE SURVEY**

## 2.1 Introduction:

It is assumed that most of the effort spent on data warehouse is in its development stage which can be attributed as back end issues, such as extraction, loading and transporting data into data warehouse. But here is another main factor that is to maintain and ensure the data warehouse quality.

There are so many quality factors in data warehouse like quality data factor that is to ensure that no dirty data is populated into data warehouse similarly quality information, quality goals and quality decisions. All these factors effects and ensures the quality of data warehouse, thus enhancing the usability of the warehouse.

In all areas of information resource management, data quality is the most critical issue. As data warehouse as a decision support system is largely based on the decisions that are made on information extracted from data stored in data warehouse.

## 2.2 Data Flow:

As shown in figure 2.1 the development steps of data warehouse or repository are well understood. In the figure 1 data flow from heterogeneous source databases into data warehouse repository after going through intermediate staging area. There may be data quality tools available to transform the data & ensure its quality which helps in enhancing the usability of the data once it reaches to the data warehouse.

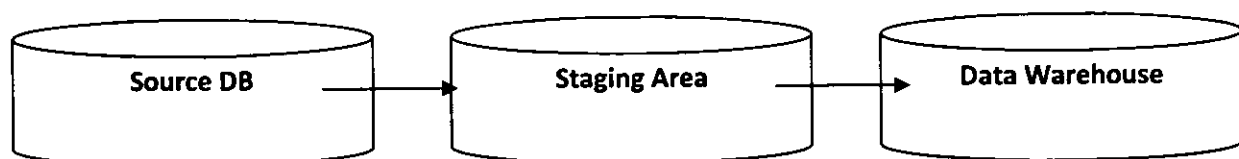


Figure 2.1: Data Flow [14]

### 2.3 Data Quality Review in Data Warehouses:

Nowadays with the importance of decision support system, data cleansing activities account largely. In this regard so many tools are made available to automate the part of the tasks which are involve in extraction, cleansing auditing and loading data into the data warehouse. Among these tools large number of tools comes in the extraction & loading of data classification and remaining small number of tools are cleansing and auditing tools.

Historically, IT personnel build few of their own techniques for cleansing data i.e. data entry based validation like what kind of data should stored in the field, it also checks for its reasonability and perform some other validation checks. Data quality tools emerged as a new way to clean & verify data at maintaining and development stages of data warehouse. Through these tools data is audited at data source and transformation level so that data become consistent throughout the warehouse. To ensure that the data match the business rules these tools can also be used to segment the data into granules or atomic unit. We can say these data quality tools are stand alone packages which can be integrate with other data warehouse features.

Different tools that are used in data flow from Source data bases to staging area and then to data warehouse repository are shown below in figure 2.2

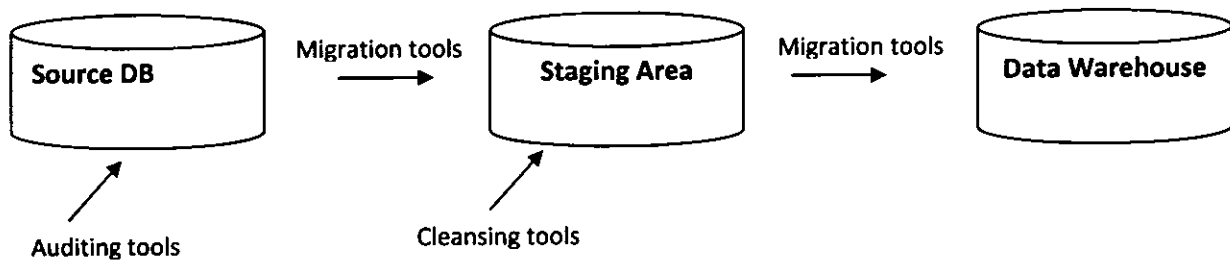


Figure 2.2: Data Flow & Data Quality tools [14]

## **2.4 Data Quality Tools:**

Data quality tools are categorized into 3 ways i.e. Auditing, cleansing and migration.

### **2.4.1 Data Auditing Tools:**

These tools use for accelerating the correctness and accuracy of the data source. The technique used in auditing tool is basically comparing the business rules with the source database. Business rules are determined by using some techniques of data mining to know the patterns of data, when organization is using some external source of data. Business rules should be apply at early stage to external data sources when they are internal to the organization.

Business rules are of vital importance to data because the data which don't stick to the business rules cannot be modified further when necessary.

### **2.4.2 Data Cleansing Tools:**

In the staging area at intermediate level of data flow this tool is used. These tools are meant for cleansing purpose like an independent source data, addresses, names etc. These tools are well known for standardizing parsing and also to verify the data against some kind of known lists (matching)

Data cleansing tools performs functions which are given below:

#### **2.4.2.1 Data Standardize:**

It converts the source data into a standard form that is following throughout the data warehouse.

E.g.; all incidences of budget should be represented like 'bdg', not budget or bud.



#### **2.4.2.2 Data Parsing:**

This tool is used for making small granules of the data record which can be used in subsequent steps further. Data Parsing also used for placing the elements of record into their correct fields.

#### **2.4.2.3 Data Correction & Verification:**

This tool is used for the purpose of verifying data to some already known list and then correcting it. It matches the data against known lists.

#### **2.4.2.4 Record Matching:**

Record matching tool is used to find out the duplicate data whether are there any records representing data on the same subject.

#### **2.4.2.5 Data Transformation:**

Data transformation tools ensure the consistent mapping between the data source system and data warehouse.

For example: '1' for female becomes 'F'

'2' for Male becomes 'M'

#### **2.4.2.6 Documenting:**

This tool creates a log file for information in meta data after clearing steps.

## **2.5 Data Migration Tools:**

This tool is used to transfer the data from the source system to the data warehouse after going through the staging area where cleansing tools are evoked. Data migration tools are responsible for converting the data from one platform to another.

## **2.6 Data Warehouse Critical Success factors:**

Critical Success factors are the factors that lead to implement an efficient Data warehouse system.

There are several factors that are crucial in the way of developing an efficient data warehouse system. These factors can be following:

1. The right architecture, proper design and implementation of data bases can greatly improve and also ensures performance today and scalability tomorrow.
2. Data warehouse components which include network, data repository, user interface and application logic etc, they all must be truly coordinated with each other in a flexible & easy to use manner.
3. Third crucial factor is to build a consistent data model which could tell and justify that how and what source data should be extracted.

## **2.7 CSF in Data Warehouse Technology:**

Following are the main critical factors that can lead to an efficient implementation of data warehouse.

- Organizational Factors
- Environmental Factors
- Project Related Factors

- Technical Factors
- Educational Factors

In Figure 2.3, research model for Data warehouse critical success factors is presented.

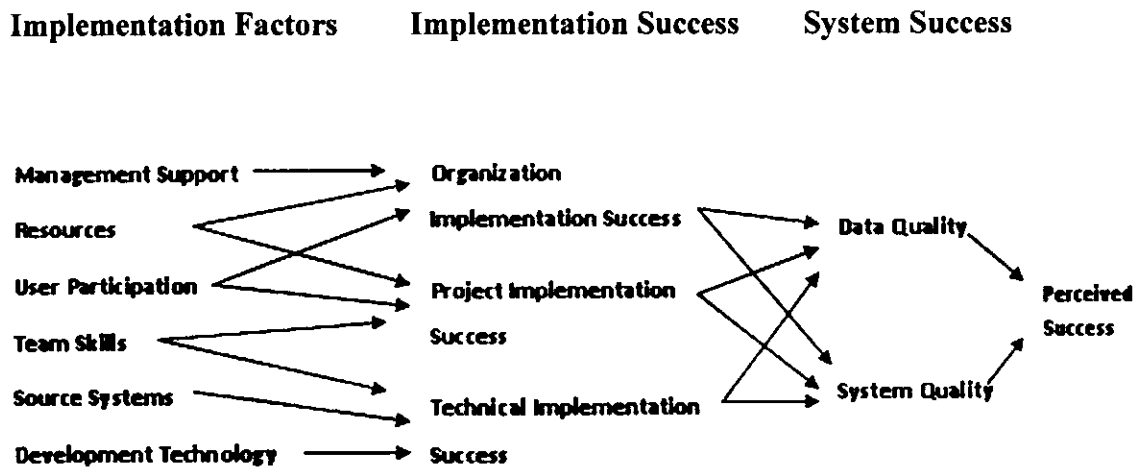


Figure 2.3: Research Model for Data warehouse Success Factors in FAC Corporation [1]

Critical Success factors may also vary due to difference in nature and work of many organizations whether is it better to adopt decisional factors for the better accomplishment of strategic business objectives or critical success factors should be implementation factors that may lead to system success for DW implementation so all these things strive for organizational success with project & technical issues during the whole life time of DW project. System quality and data quality greatly influenced by these successes.

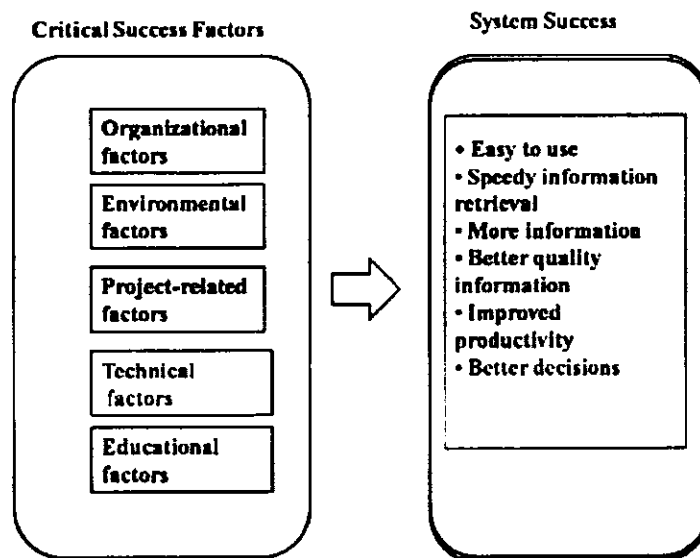


Figure 2.3.1: Data warehouse CSF [1]

### 2.7.1 Organizational Factors:

Data warehouse critical success factors with the dimensions of organization are given below :

- Size of the organization
- Existence of champions
- Executive Support
- Business requirement and interval needs
- Organizational resistance
- Organizational policies

#### 2.7.1.1 Size of the Organization:

An adoption of data warehouse technology should be with respect to the size of the organization. Normally big organizations have more resources and capital to invest on a Data Warehouse project.

### **2.7.1.2 Existence of Champions:**

Brilliant minds inside the organization always use to support and appreciate the adoption of new technology, and also encourage the staff to adopt the new techniques & technologies.

### **2.7.1.3 Executive Support:**

Executive Support is very important in some complexities which arise in data warehouse project. This Support can greatly secure required capital, availability and coordination between internal resources, human resource support etc.

### **2.7.1.4 Business Requirement and Internal Needs:**

To apply business techniques on data warehouse project is also a very crucial step in its development. Business requirement and internal needs becomes the basis for the organizational information architecture and as well as data warehouse architecture and design.

### **2.7.1.5 Organizational Resistance:**

Employee's resistance inside organization is basically due to the fear of losing their jobs by replacing labor work with automation after the implementation of the new technology. Such kind of resistance should be handled appropriately by training the staff and encouraging them to adapt the new technology.

### **2.7.1.6 Organizational Policies:**

Every organization has their own rule & regulations to govern and control the organizational activities & processes to achieve the long term goal and strategic objectives. In data warehouse matter it is strongly recommended to align the data warehouse technology. In these concern organization policies provides

the basis in the form of detailed information about aligning the data warehouse technology to achieve the required goals.

### **2.7.2 Environmental Factors:**

The Enterprises having dynamic nature of environment mostly have high ratio of sudden changes. So enterprises must adopt new technologies to track the changes and to stay competitive in the surrounding environment.

Following is the list of environmental factors. [1]

- Business Competition
- Selection of Vendors
- Governmental regulations & Compatibility with industry
- Compatibility with Partners.

### **2.7.3 Project Related Factors:**

Project related factors in the favor of adopting data warehouse technology are of utmost importance. Project related factors can be many like Project Plan, Development & Control, and Analysis etc [5]

Following are the project related factors:-

- Skills of Project Team
- Emergence & Coordinating Organizational resources.
- End-user participation or Involvement.
- Support from Experts.
- Accurate definitions of Project Priorities, scope & goals

### **2.7.4 Technical Factors**

During data warehouse implementation stage some technical problems arise and these problems comes in the technical factors which should be considered at every stage of Data warehouse development.

Sub functions under technical dimensions are:

- User interface
- Technical resources availability & proper technology development
- Quality of data sources.

### **2.7.5 Educational Factors**

Educational factors are about the training of the existing human resource and encourage them to adopt new technology. As this can lead to enhance the interaction b/w employees & new technology and also leads to widen the knowledge of new technology. [6]

Educational factors can be:-

- Training Courses
- Certified Trainers
- Availability of best practices adaptors.

All the above factors considered very important in the concern of data warehouse architecture & design. The architecture & design shows the requirements of the enterprise and reflects the performance measurement.[7] Data warehouse data model, Meta data structure components should be made on the basis of internal information requirements. [8]

All the CSF improves the quality of data warehouse in terms of:-

- Easy to use
- Speed retrieval of information
- More information
- Better quality information
- Improved productivity
- Better decisions

## **2.8 META-DATA REPOSITORY:**

In data bases, Meta data repository term is used for the information of data or data about data. Meta data repository is used for the purpose of tracking all the records and changes stored and made in data bases, highly reliable and consistent way to have access of the information stored in data ware houses. Meta data repository may be stored physically or logically in data warehouses, from which Meta data is drawn from separate multiple data sources. Metadata can also include information about how to get access of specific data from huge data warehouse and to get more detail about it, among bundle of possibilities.[9]

Metadata is like a concept that is used mainly for archived data like in databases or data warehouses and is used to describe the following:-

a) Definition

b) Structure

c) Administration of data files with all contents in context to ease the use of the captured and archived data for further use.



### 2.8.1 Meta-Data types:

There are many well accepted models to specify types of meta-data.

Bretheron & Singley in 1994 differentiated between two distinct classes i.e. ,

- Structural / Control meta-data
- Guide meta-data

#### 2.8.1.1 Structural meta-data:

Structural meta-data is basically use to elaborate and describe the structure of database systems such as tables, indexes and columns.

#### 2.8.1.2 Guide meta-data:

Guide meta-data used for the purpose of a guide who helps users to find specific items and normally known as a set of keywords.

According to Ralph Kimball meta-data can be divided into categories which is :-

- Technical meta-data
- Business meta-data

Technical Meta data parallel to internal meta-data where as Business meta-data parallel to external Meta data.

NISO differentiate between three types of meta-data i.e Descriptive, Structural and administrative. Descriptive is to describe the information used to search objects such as subjects, keywords, title, author

and publisher etc. Structural describes the organization of the components of the objects. Administrative meta-data involves the technical information of the database which also include file type.

### **2.8.2 Meta-Data and Data Warehouse:**

Data warehouse (DW) is a kind of repository in which organizations stores data electronically. Data warehouses are made to store & manage huge amount of data whereas Business Intelligence (BI) gives main focus on how to use the stored data to facilitate the analysis and reporting.[10]

The purpose of a data warehouse is to house standardized, structured, consistent, integrated, correct, cleansed and timely data, extracted from various operational systems in an organization. The extracted data is integrated in the data warehouse environment in order to provide an enterprise wide perspective, one version of the truth. Data is structured in a way to specifically address the reporting and analytic requirements.

An essential component of a data warehouse/business intelligence system is the metadata and tools to manage and retrieve metadata. Ralph Kimball describes metadata as the DNA of the data warehouse as metadata defines the elements of the data warehouse and how they work together. [10]

## 2.9 Related Work:

### Meta-data repository & Quality Models:

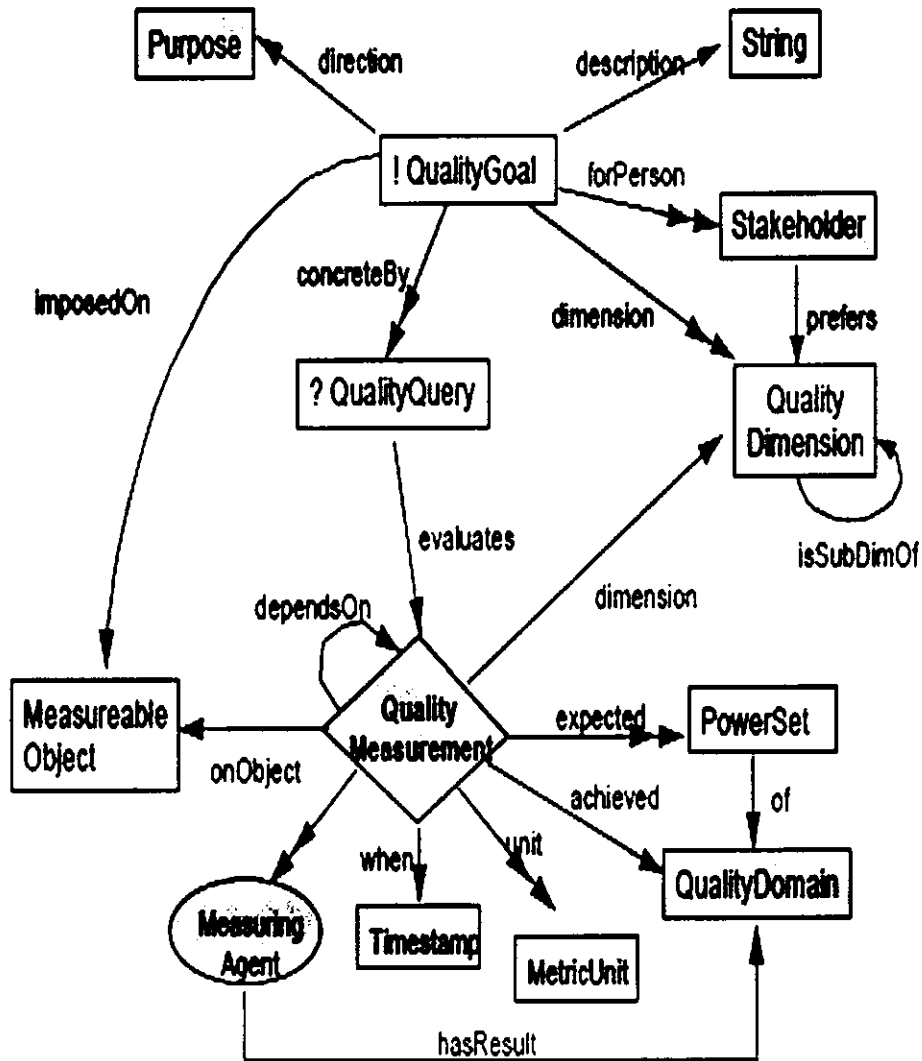


Figure 2.4 Goal Question Matrix (GQM) Approach [02]

In this model Goal Question Matrix (GQM) approach is adapted.

- Requirements are modelled as goals.

- Assist users to create & evolve the knowledge base their views on the federated database system.

A limitation of this model is that when & how the views are populated and updated.

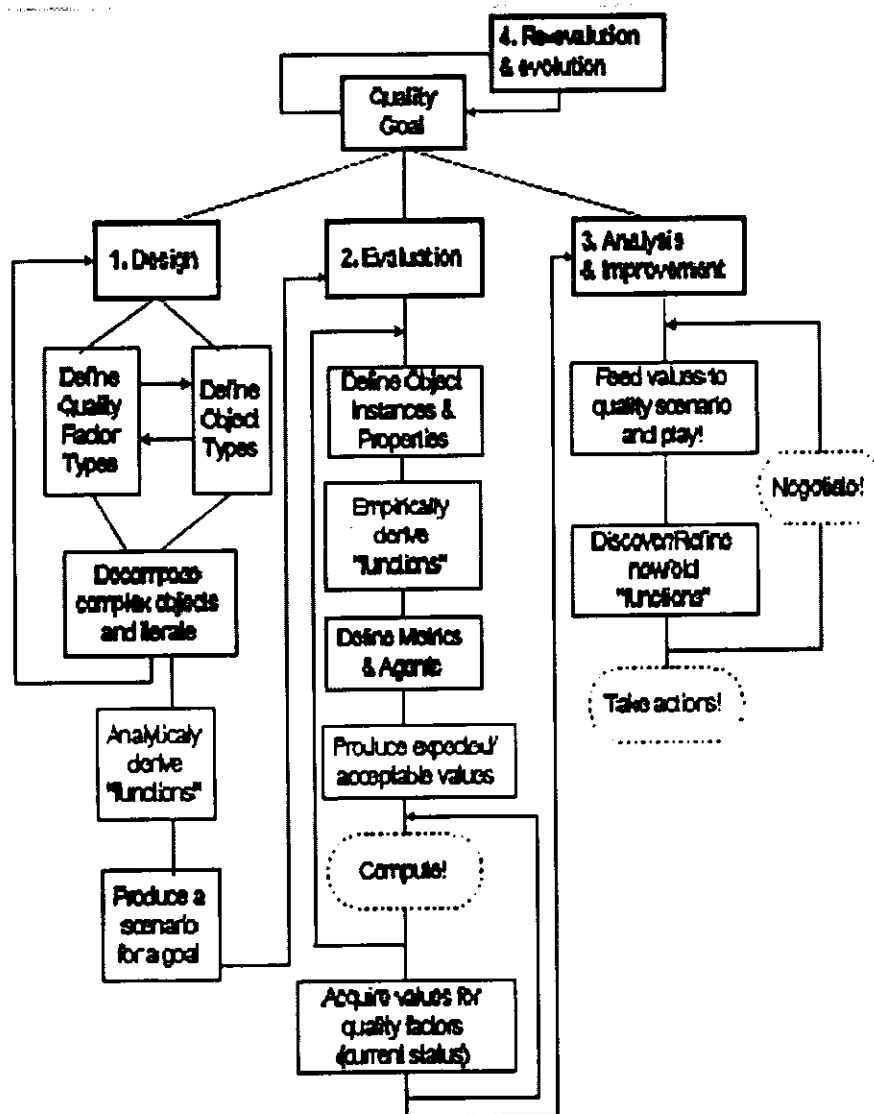


Figure 2.5 Extended Goal Question Metric Approach (GDI) [17]

This model is an extension of the Goal-Question-Metric (GQM) Approach, which allows to capture:

- The inter-relationships between different quality factors &
- To organize them in order to fulfil specific quality goals.

Limitations of the (GQM) model is due to interdependencies and computation fractions between quality factors.

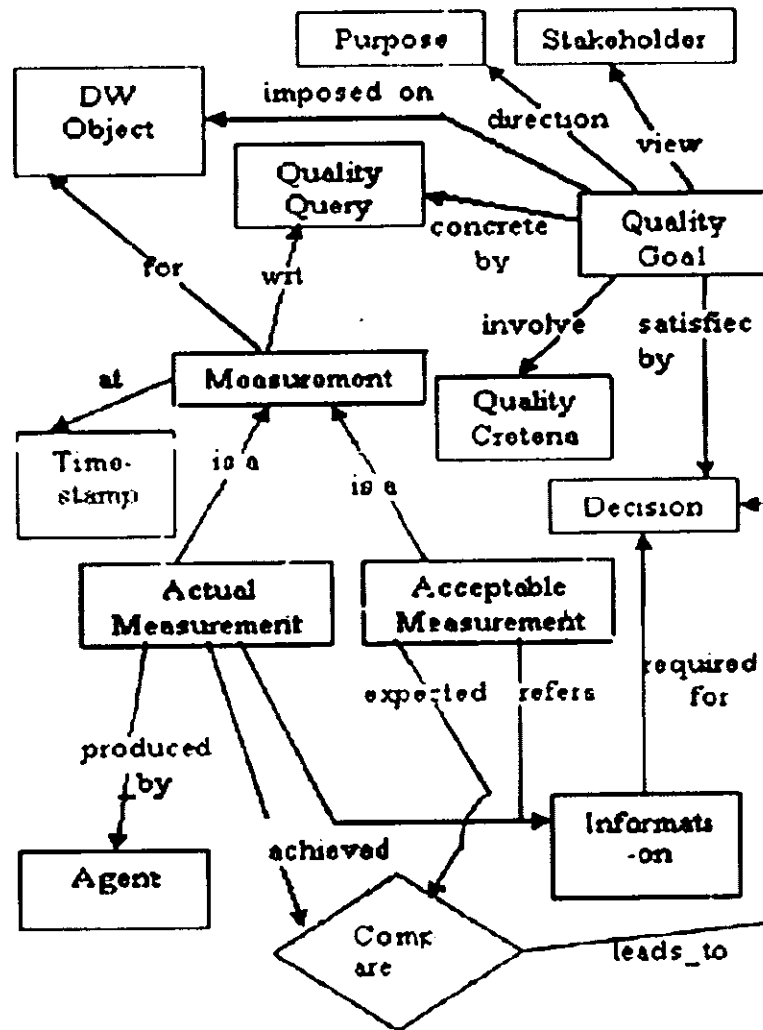


Figure 2.6 Goal Decision Information (GDI) Approach [9]

- GDI approach is a goal oriented model in which basic aim and objective is a 'Goal' which has to be achieved.
- A goal is a passive concept. It's not like an event, activity or process which are related to some kind of action.
- When a goal has been set, then to meet the goal some active component like decision is required.
- Decision needs an appropriate, up to date and correct information for the fulfillment of the goal.

## 2.9.1 Literature Survey:

### 1. Extension of the Goal-Decision-Information Approach to Ensure Quality in Data Warehouse Development Using Software Agent

The seminal study in this area is by S.VenkatesanI, S.Mohamed Saleem, in 2009 who argues in the Extension of the Goal-Decision-Information Approach to Ensure Quality in Data Warehouse Development Using Software Agent that Data warehouses are complex systems that have to deliver highly-aggregated data from heterogeneous sources to decision makers. It is essential to assure the quality data warehouse in terms of data as well as the services provided by it. But the requirements and the environment of data warehouse systems is dynamic in nature. To handle these changes efficiently, data warehouses depend largely on the Meta databases. He has extended the Goal-Decision-Information approach to model the quality of the data warehouse. In order to fulfill the specific quality goals, dependencies among the various quality factors is exploited in this model. [6]

#### **Existing Technology used:**

Existing model that was presented in 2009 was extension of Goal decision quality model which achieved performance up to the some extend but common issue related to it is there is no suitable strategy given for materializing the quality measurements which is missing and also measuring the quality can be computationally expensive.

The idea of GDI is that quality goals can usually not be met directly, but their realization is circumscribed by decisions that need to be taken, making use of the information available. Such decisions again can usually not be taken directly but rely on metrics applied to either the product or the process which relates to the goal in question; specific techniques and algorithms are then applied to derive the answer from the measurements.

**Limitations:**

1. A suitable strategy for materializing the quality measurements is missing. For the moment, we assume that there are external metric agents that compute some quality value for a given measurable object. But when should the agent be activated? Supposedly, measuring the quality of a component like a source relation is computationally expensive.
2. A suitable collection of quality metrics for data warehouses has to be investigated. Starting point is simply cannot afford to measure the quality.
3. To extend to method to the design of a data warehouse which includes selection of the right source databases, filters, transport agents etc. based on their quality properties.

**2. The Research in Improving the Quality of DW Data**

A Similar study done by Jie Zhang, Qiaoyan Wen, Hua Zhang in 2009 is "The Research in improving the quality of DW data". There are plenty of sophisticated database application systems in telecommunications industry, such as "integrated management system based on telecommunication services", "Local Network Management System", "Financial management system" and etc. These systems generated a large amount of business-processing data. However, much of the historical data have been stored on tape, CD-ROM, or in different hardware and databases. So in comparison of the data queried and analyzed by different business units, there will be many problems such as the mismatch of types, inconsistent data, data redundancy and etc. In order to solve these problems, the data warehouse technology is worked out. This paper will introduce the design of the ETL (Extract, Transform, Loading) system, job scheduling and the checking of calibrating process through an analysis of the actual system in EMC(ETL Manage Center) products. [7]

**Existing Technology used:**

The detailed analysis for ETL framework in EMC Products is completed by using software project ideology. And the data scheduling and checking process on background is focused on. The practicality and rationality of this process in EMC Products is

Verified according to the cases.

**Limitations:**

This work is more suitable for centralized databases, and this technique is not practically implemented for heterogeneous data bases.

### **3. Towards Data Warehouse Business Quality through**

#### **Requirements Elicitation**

There is another approach adopted by Anjana Gosain, Jaspreeti Singh in 2008 is “Towards Data Warehouse business quality through requirements elicitation.”

Data warehouses are mainly used to support decision-making based on the analysis of highly heterogeneous sources to extract, transform and aggregate data, as well as facilitating ad-hoc queries that retrieve the decisional information. Data warehouse development involves many knowledge-intensive activities, of which requirements elicitation is recognized as being crucial and difficult to model. This paper adapts the data warehouse requirements elicitation process, namely Informational Scenarios, to incorporate business quality at the requirements engineering level of the DW development. To accomplish this DW business quality mainly from the context of changing economic factors and environmental concerns. [8]

**Existing Technology used:**

Informational scenarios elicit the informational requirements of decisional systems. In this paper, integrating the concepts of requirements engineering and business quality, and is looked at informational



scenarios from the point of view of business quality, providing a wider solution to the requirements elicitation process for data warehouse systems. The interest is mainly in obtaining information keeping into consideration the aspects which contribute in achieving business quality of the system. Based on the availability/ unavailability of the information, the decision maker may formulate sequence of interactions to reveal other information.

**Limitations:**

The limitation is to provide complete framework for integration of quality at the requirements engineering level, which is a future research question.

#### **4. An Application of Data Mining to Identify Data**

##### **Quality Problems**

In an application of data mining to identify data quality problems by Eshref Januzaj, Visar Januzaj in 2009 a much similar study is done.

Modern information systems consist of many distributed computer and database systems. The integration of such distributed data into a single data warehouse system is confronted with the well known problem of low data quality. In this paper an approach that facilitates a dynamic identification of spurious and error-prone data stored in a large data warehouse is presented. The identification of data quality problems is based on data mining techniques, such as clustering, subspace clustering and classification. [9]

**Existing Technology used:**

An approach is presented to identify data quality problems by applying data mining techniques (clustering, subspace clustering, and classification). A density based clustering algorithm to generate data object clusters. The resulting clusters were then used to build the data classifier, which in turn is used for the identification of low quality data in the entire data warehouse. This approach has been tested on an applied to real data.

ODS's, which would not be used to identify with common analysis tool. Other data quality problems are as inconsistencies, redundancies.

## **5. Methodology for Information Quality Assessment in Data Warehousing**

A similar study was conducted by Ying Su, Zhanming Jin in 2008 who argues a methodology to determine two IQ characteristics—accuracy and comprehensiveness—that are of critical importance to data warehousing. This methodology can examine how the quality metrics of source information affect the quality for information outputs produced using the relational algebra operations selection, projection, and Cubic product. It can be used to determine how quality characteristics associated with diverse data sources affect the quality of the derived data. The study resulted in the development of a model of a data cube and an algebra to support IQ Assessment operations on this cube. [10]

### **Existing Technology used:**

An important issue within the realm of decision support databases is addressed: the lack of a precise, commonly agreed upon conceptual model for assuring the quality of information. To address this problem, two significant contributions are made. First, a detailed data model is presented for the data cube. Secondly, a detailed operations model for the data cube to determine how source data cube of different quality could impact those OLAP derived using Selection, Projection, and Cubic product operations is presented. Models can be used in several ways.

For example, consider selecting prospective customers for a promotion using in-house customer transaction data along with geographical data purchased from an external vendor. Data from the two tables would need to be joined and then the appropriate selection condition applied to the result of the join. The join requires a Cartesian product operation that would typically lead to an increase in the mismembership of the resulting table as compared to either of the participating relations. The selection operation would further increase the mismembership in the target address tag.

The estimates for incompleteness provided by our models would help determine whether additional data should be purchased from vendors. Because data mining could support multiple such applications, our analysis would be useful in identifying which data sets will have acceptable quality, and which ones will not. Finally, our results

can be implemented on top of data warehouses engine that can assist end users to obtain quality profiles of the information they receive. The quality information will allow users to account for the reliability of the information received thereby leading to decisions.

## **2.11 SUMMARY:**

In this chapter we have seen different issues related to 'quality of data' and data flow. We have gone through the different data quality tools and its impact on the quality of data warehouse.

Data warehouse critical success factors are also mentioned which leads to implement an efficient data warehouse system, each factor is discussed briefly. Then the importance of Meta-data repository is briefly touched and explained some of its types.

At the end, the previous architectures and models developed to improve the data warehouse quality are mentioned and their significance and limitations are briefly discussed.

**CHAPTER NO. 3**  
**REQUIREMENT ANALYSIS**

### 3.1 Introduction:

Nowadays data warehouses have become very complex systems which are very useful for decision makers by analyzing the highly aggregated data from heterogeneous sources. Due to the importance of information which is extracted on the basis of data stored in data warehouses it became very essential to ensure the quality of the data stored in it, and also quality of overall data warehouse which can be achieved by improving the quality of data and services provided by it. [13]

But this is not such an easy task especially when it comes to dynamic requirements and dynamic kind of environment of data warehouse systems in nature. To handle all these issues effectively & efficiently data warehouses largely rely on the Meta databases.

A data store of low granularity and volatile data is used mostly for the integration of data obtained from the various data sources which is commonly known as (ODS) Operational data store.[8] At ODS level the cleaning processes and data transformations are carried out to ensure the clean and harmonized data populated into data warehouse. At the top layer of data warehouse architecture highly aggregated data is present which is obtained from the global warehouse.

So many components make the data warehouse systems, which may involve huge no of stakeholders having different objectives & goals. These goals are monitored constantly through tools like administrative tools. The most important thing is that all the data warehouse processes, involving many components and data is tracked through the metadata.

The data warehouse where holding large amount of data makes it a very complex system which involves multiple processes like aggregation of data, extraction, cleansing, transformation and storage etc. For the design choices and record of changes, the metadata repository is like a path to keep track the changes performed on its components and architecture. [3]

### 3.1.1 Meta-Data Repository:

In data bases, Meta data repository term is used for the information of data or data about data. Meta data repository is used for the purpose of tracking all the records and changes stored and made in data bases, highly reliable and consistent way to have access of the information stored in data ware houses. Meta data repository may be stored physically or logically in data warehouses, from which meta data is drawn from separate multiple data sources. Metadata can also include information about how to get access of specific data from huge data warehouse and to get more detail about it, among bundle of possibilities. [2]

Metadata is like a concept that is used mainly for archived data like in databases or data warehouses and is used to describe the following:-

- a) Definition
- b) Structure
- c) Administration of data files with all contents in context to ease the use of the captured and archived data for further use.

### 3.1.2 Meta-Data types:

There are many well accepted models to specify types of meta-data.

Bretheron & Singley in 1994 differentiated between two distinct classes i.e.

- Structural / Control meta-data
- Guide meta-data

**Structural meta-data:**

Structural meta-data is basically use to elaborate and describe the structure of database systems such as tables, indexes and columns.

**Guide meta-data:**

Guide meta-data used for the purpose of a guide who helps users to find specific items normally known as a set of keywords.

According to Ralph Kimball meta-data can be divided into categories which are:-

- Technical meta-data
- Business meta-data

Technical Meta data parallel to internal meta-data where as Business meta-data parallel to external Meta data.

NISO differentiate between three types of meta-data i.e. Descriptive, Structural and administrative. Descriptive is to describe the information used to search objects such as subjects, keywords, title, author and publisher etc. Structural describes the organization of the components of the objects. Administrative meta-data involves the technical information of the database which also includes file type.

**3.2 Meta-Data and Data Warehouse:**

Data warehouse (DW) is a kind of repository in which organizations stores data electronically. Data warehouses are made to store & manage huge amount of data whereas Business Intelligence (BI) gives main focus on how to use the stored data to facilitate the analysis and reporting.

The purpose of a data warehouse is to house standardized, structured, consistent, integrated, correct, cleansed and timely data, extracted from various operational systems in an organization. The extracted data is integrated in the data warehouse environment in order to provide an enterprise wide perspective, one version of the truth. Data is structured in a way to specifically address the reporting and analytic requirements.

An essential component of a data warehouse/business intelligence system is the metadata and tools to manage and retrieve metadata. Ralph Kimball describes metadata as the DNA of the data warehouse as metadata defines the elements of the data warehouse and how they work together.

### **3.2 Critical Scenarios:**

In recent computer age, the focus is given to achieve both high quality services as well as costs down. In order to make effective and efficient decisions there is great need to enhance the quality of data warehouse system.

Data warehouse is specifically intended to provide the decision makers with the vital strategic information that's why data warehousing is like a new paradigm. The top executives and managers need proper information so that it could be helpful in making proper decisions to keep their enterprise competitive, basically for organization, the information is critical to formulate the business kind of strategies, set organizational objectives, establish goals and then monitor results.

To make decisions in the execution and making business strategies & objectives, the kind of information that is needed is broad based and encompasses all the information of the whole organization to be included in one group and that is called strategic information.



The different approaches that are followed in data warehouse development to support decision making activities in an organization are data-driven and requirement driven.

In data driven data is collected from different operational systems into the data warehouse. Whereas in requirement driven approaches, try to identify the necessary information that needs to be met by the requirement of the data warehouse.

An ideal quality data warehouse system would encompass all possible and relevant information which should be accurate and every possible alternative. Time and effort are the two constraints in extracting information and to identify the limited alternatives. The time constraints says that up to a certain limit of time, a certain decision must be made, where as the effort constraint focuses on money, priorities and reflects the limit of man power. So to enhance the quality factors all these issues are critical to data warehouse development.

### **3.3 Focus of Research:-**

People usually think that Quality of data warehouse term is related to the quality of data, however quality with respect to data is very important but in fact the term quality refers to a bigger picture.

According to David Wells and James Thomann;

“Quality is the act of measuring the progress of the data warehouse in terms of its ability to satisfy your customer base”.

Data quality is not enough- decisions are made on information quality and not only on data quality and data is only fit to be treated as a referenced for data quality measurements.

In this research work, my focus is on quality of the data warehouse system. Quality of Data warehousing can be checked directly through Meta data repository. Current data ware Meta models can't express large number of quality factors relevant for data warehousing. So in this research focus is given to enrich the Meta data model of data ware architectures by exploiting the techniques for measuring or optimizing the specific aspects of data warehouse quality factors.

Data warehouse as a decision information system must provide accurate and relevant information to help & support in making a specific decision. Various quality models that have been presented shows how semantically rich meta-information can be stored in a meta-data repository of a data warehouse. In this regard different quality factors are exploited regarding data warehouse components and their linkages by an operational methodology, to show how these quality goals can be achieved to enhance the quality of the data warehouse.

### 3.4 Summary:

In this chapter a brief introduction is given in the start then meta-data repository and its types are discussed briefly. Short description of meta-data and its importance in data warehouse is given.

At the end of the chapter critical scenario and focus of research is elaborated.

**CHAPTER NO. 4**

**PROPOSED ARCHITECTURE**

## 4.1 Introduction:

As we know, high aggregation of data that comes from multiple heterogeneous data sources has increased the complexity of data warehouse system and makes it more complex system. In this case it became very important to make sure that data warehouse is consistent and also to assure its quality.

Data warehouse quality can be judged in perspective of data as well as the services provided by it, but due to involvement of dynamic nature of data warehouse system, these changes cannot be handling easily. In this scenario data warehouse system largely depends on the Meta databases to deal with the changes efficiently which are brought by the dynamic nature of data warehouse.

Quite a lot of research has been done and different Architectures and models have been proposed to improve the efficiency and quality of data warehouse system. Goal Decision information (GDI) approach is one of them. Here the proposal is to extend the existing GDI model to improve the quality of data warehouse system and remove the deficiencies left in the existing model.

The idea of Goal Decision Information (GDI) is that sometimes quality goal cannot be reached or achieved directly and they are restricted to the decisions that have to be taken; in simple words quality goals are highly dependent on the decisions. Decisions can be taken by making use of the available information in the data warehouse system. Such decisions indirectly depend on the processes and products and rely on the metrics which are applied either to the product or processes, which usually relates to the goal in question.

In the proposed approach, a criterion has been set which is achieved through a strategy derived from quality factors, and a strategy is used to set the quality goal through the stakeholder or quality manager. This strategy will be based on the organizational quality factors.

## 4.2 Design Requirement:

Design requirement involve stage by stage philosophy to ensure that you are proceeding in a systematic way.

The idea of quality model is that the whole process of setting quality goal and implementing strategy is broken into different separate stages. At the level of stakeholder (quality manager), quality attributes for the data warehouse is accessed and figured out so that the dependencies between these attributes could be removed at conflicting stage.

The next stage at prioritization level evaluation of different quality attribute on the basis of priority is set. Quality goal can be set after realization the priority of the quality attribute.

A key point is that the design from one stage of the process becomes the requirements for lower stages for example, a mission statement having 5 point of strategy; these strategic points are the design that results from the mission statement requirement. Then these points become the stages of the Architecture.

## 4.3 Reference Model:

Goal Decision Information approach is a goal oriented model in which basic aim and objective is a 'Goal' which has to be achieved. A goal is a passive concept and it's not like an event or any kind of activity or process which are related to some kind of action, which can be performed. And when a goal has been set then to meet the goal and to attain it some active component name decision is required. For the decision to be taken, the very important thing is information usually stored in the data warehouse in the form of data from which we extract useful information. So decision needs an appropriate, up to date and correct information for the fulfillment of the goal.

GDI model revolves around the two kinds of goals which can be simple goal or complex goal. Simple goal can further be divided into simple ones. Complex goal may be itself derived from simple or complex ones.

Decision makes the action to be performed which makes the goal to be implemented. Decision itself is not an active component it's just a specification of active component which then use to achieve the goal. When a decision is made then one or maybe more actions are to be performed to give effect to it.

Information is used in decision making activity. Selection of right decision is highly dependent on the right information, which gathered from the data stored in the data warehouse system. A complex decision can be decomposed into simpler ones or into complex ones but simple one cannot be further decomposed.

In the figure below a connection between goal and decision is shown named as 'is satisfied by'. This association means that Goal is satisfied by the decision.

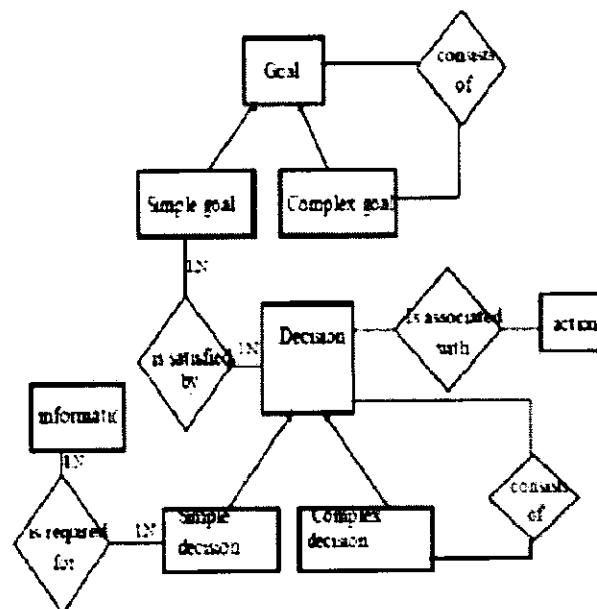


Fig 4.3: The Goal-Decision Information Model [2]

Another relation ‘is required for’ is used between the decision and information. This association tells that to take a decision information is required. The data stored in the data warehouse will specify the information eventually.

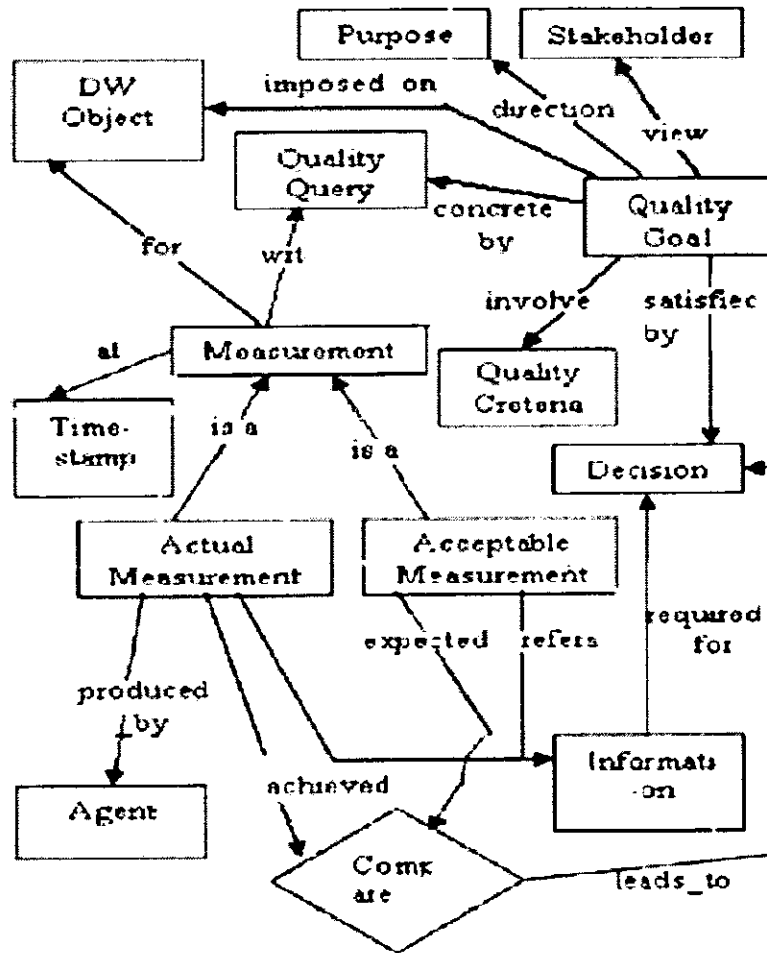


Figure 4.3.1: GDI Quality model [2]



#### 4.4 The Proposed Model:

The Proposed quality model adopts three layers of instantiation which derived from the model mentioned in the previous section, a model which logically supplements the Architecture which consists of three levels.

At the top level generic frame work of Goal is given, along with data associations where applicable. In the next level Quality Goals are specifies and in the third level of Goal Decision Information (GDI) model concrete values of measurement are mentioned.

Basically GDI is a project in which stakeholder is either a part of it or has to manage the quality of data warehouse. In the data warehouse administrator perspective, this roughly defines natural language requirements like 'achieve the availability of source S1 at least once per week.

The purpose of Goal involves quality criteria which usually concrete by different quality requirements. These requirements are used to evaluate quality factors of the organization, defines on data ware object. Basically quality criteria define abstractly as different quality aspects, as the stakeholder perceives it.

The concrete measurements for the quality questions are carried by making use of stored information. Measurement can be actual or acceptable and their comparison leads to the specific decision. This decision sets the ground to perform the required action for that decision.

The goal decision model assumes that the stakeholder provide the stored acceptable values in the metadata repository.

Below is the extended model of GDI approach in which quality criteria is evaluated by many quality factors of the organization defined on data ware objects. Quality criteria help in giving direction to the quality goal.

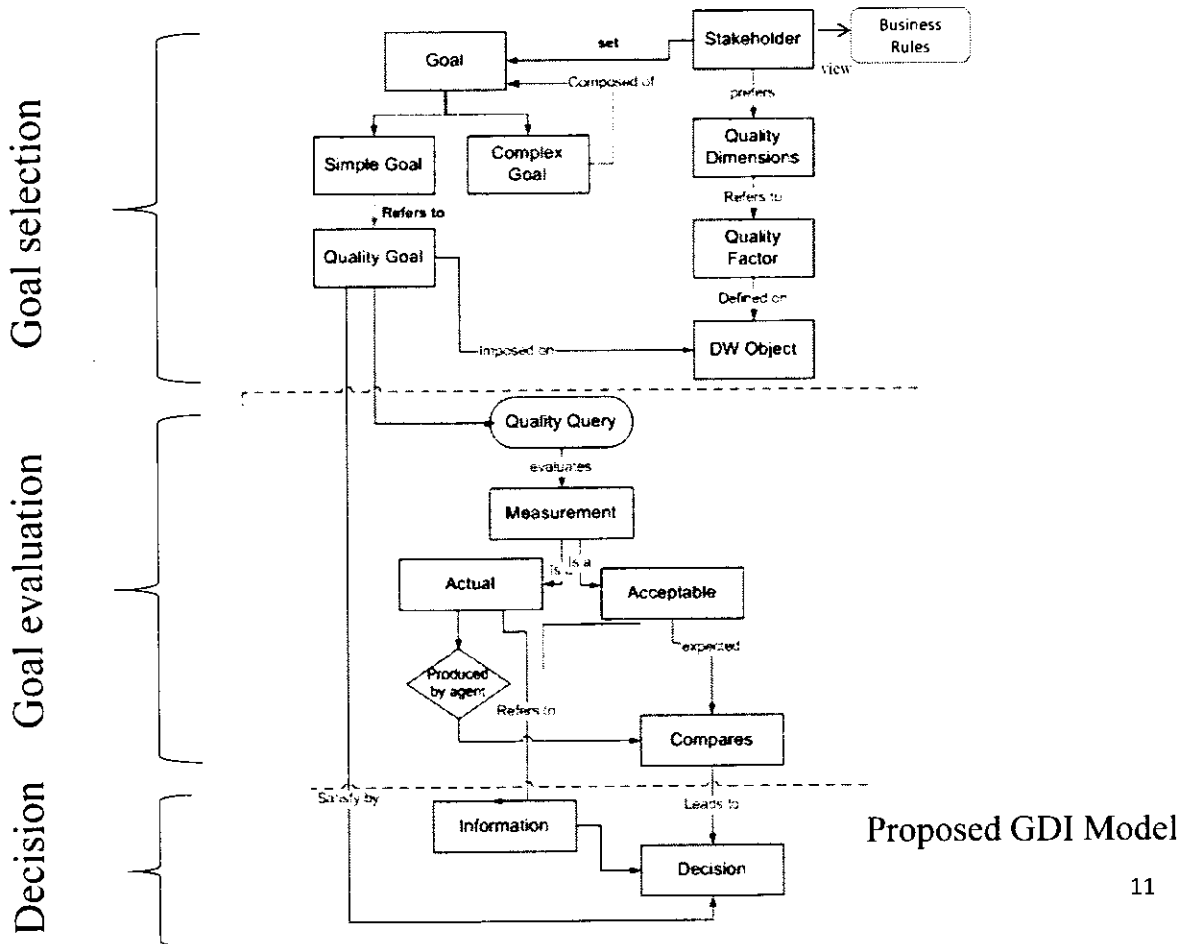


Figure 4.4: The Proposed Quality Model using GDI approach

Formally;

**Goal:** To evaluate and improve the quality of data warehouse, a stakeholder has to manage the project of goal. A quality goal decomposed into sub-goals which are defined by the following assertions:

- It refers to a data ware object.

- Possible instances of purpose type have given the direction.
- It has a reference among the instances of quality criteria entity.
- It is normally defined with respect to the specific view point of a given stakeholder.
- A goal can be refined to several quality queries.

**Purpose:** Purpose can be any reason for any action to take, to get a specific goal.

E.g.; improve, optimize, enforce etc.

**Stakeholder:** Stakeholder can be administrator analyst, designer or a person who is involved in the data warehouse project.

**Quality Criteria:** Quality criteria involve different aspects of quality. It is a subjective, high level, user oriented characterization of a given object.

**Data ware Object:** Objects are the instances of data ware house frame work, it could be any conceptual, physical, or logical at any level like at source level, client level, or enterprise level.

**Quality Query:** quality query is used for quality measurement, it is an intermediate level placed between quality goal and quality measurement which is used to mediate between the quality goal (an abstract requirement that can't be accessed directly) and a measurement (yielding concrete quality value).

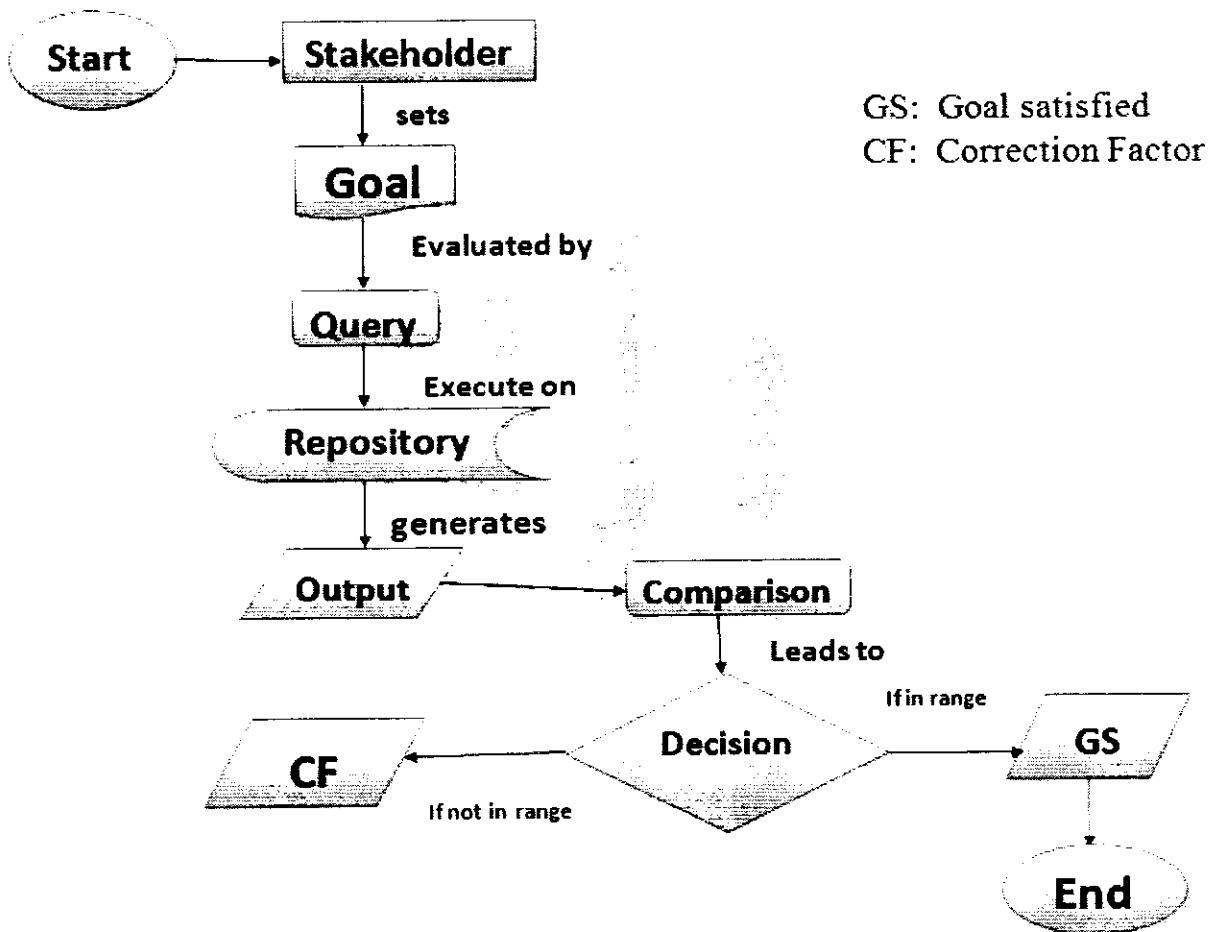
**Measurement:** measurement is used to evaluate the quality gone and is done for specific data ware object, to evaluate the quality goal at any specific time (time stamp).

**Actual Measurement:** some real measurements about quantification of the quality of the data warehouse have been formed. Measurements which represent facts is done by usually certain software program which is used for the computations and generating a specific value.

**Expected Measurement:** Expected measurement defines the interval of allowed values of the actual measurement. This interval must have the same domain with the produced values.

**Agent:** a software program of the architecture model. Each agent is characterized by a description for its functionality.

### 4.5 Flow Chart



#### 4.6 Pseudo Code:

The Pseudo Code of Enhanced Goal Decision Information (GDI)

Begin:  $G_{set}$

$G_{set}$  is set of goals set by stake holder

Process:

$G_{fset} = \text{Transformation}(G_{set})$

Actual Measurement = Query( $G_{fset}$ )

IF Actual Measurement = Expected Measurement

Then  $G_{set}$  is satisfactory

Else

CorrectionFactor (Actual Measurement)

**Output: Quality Goal is Satisfied**

## 4.7 Algorithm

The Algorithm of Enhanced Goal Decision Information (GDI)

**Begin**

Quality Goal G

Quality Factor QF(X,Y,Z)

    G ∈ QF

    Query G(X)

        Measurement (Actual Measurement, Expected Measurement)

        M in (AM, EM)

            EM = Range(0,1);

            if AM in EM

                then Quality Goal Satisfied

            else

                Correction Factor;

        End

Else

end;

**End;**

## 4.5 Summary

In this chapter, a new architecture / framework of Goal Decision Information has been proposed. It will be capable of selecting the right goal based on the requirements and keeping in view the critical factors of the organization.

The selected goal then becomes the quality goal to test the whole scenario on which decision has to be made. In this chapter previous GDI model has also been discussed and its limitations. In the next chapter implementation support of the model is given and testing scenario of single quality goal is instantiated as a case study.

## **CHAPTER NO 5**

### **INSTANTIATION OF MODEL**



## 5.1 Implementation Support for Extended Quality Model:

Quality queries are defined over quality measurement, and quality Goals are made operational through these queries. As quality queries supports to evaluate the specific quality goal, the time it parameterized with part of a Meta database.

In order to achieve the quality level, such queries mostly compare the goal analysis with certain expected interval.

Basically there are two main basic issues to be resolved, as we must organize quality goals with respect to the stakeholder analysis of the organizational requirements and quality factors because quality is a subjective phenomenon. Secondly due to dynamic nature of enterprise, executive's requirements and perspectives to stay in the competition changes with respect to change in time so stakeholder has to keep an eye on the critical factors of organization and has to define and re-define the quality goals.

The problem of introducing model for quality in the meta-data is therefore to achieve breadth of coverage without giving up the detailed knowledge for certain criteria, this can greatly help in systematic management.

On the basis of analysis, stakeholder entails multiple collections of quality dimensions and quality factors which a quality model should be address in a meaningful and consistent way.

## 5.2 Defining and Redefining Goals:

Quality goals are the 'requirements' which has to be evaluating through the quality query. We can further define the quality goal into simple & complex goals. Simple & Complex goals can be derived from user's simple & complex requirements;

For Example:

**Complex Goal:** User requirement is to quantify believability of the information delivered to the end user, here believability is the complex user requirement ultimately complex goal, which can further broke down into simple goals like into completeness, accuracy and consistency etc.

Following table can be the best illustration of simple and complex goal;

Complex Goals	Simple Goals
<b>Believability</b>	<ul style="list-style-type: none"> <li>• Accuracy</li> <li>• Completeness</li> <li>• Consistency</li> </ul>
<b>Usefulness</b>	<ul style="list-style-type: none"> <li>• Relevancy to the DW</li> <li>• Timeliness</li> <li>• Data usage</li> </ul>

Table 1: The Goal scenario

Where timeliness again a complex requirement which can further achieve through Source concurrency, DW concurrency or non-volatility.

<b>Timeliness</b>	<ul style="list-style-type: none"> <li>• Source concurrency</li> <li>• DW concurrency</li> <li>• Non-volatility</li> </ul>
-------------------	--

### 5.3 Quality Measurements:

Quality measurements can be achieved as explicit relationships between the abstract representation between the objects that are measurable and the quality values.

**For example;** we want to measure the percentage of null values for a specific relation of some data source. Let's say relation 'department'. Suppose 0.8 is the value measured for the percentage of null values. So the quality measurement recorded for the table departments is 0.8 which is just a number.

As a consequence, the measurement of quality must contain information for the actual & expected values. So they could be recorded into the meta database manually or computed through an agent or some specific reasoning mechanism.

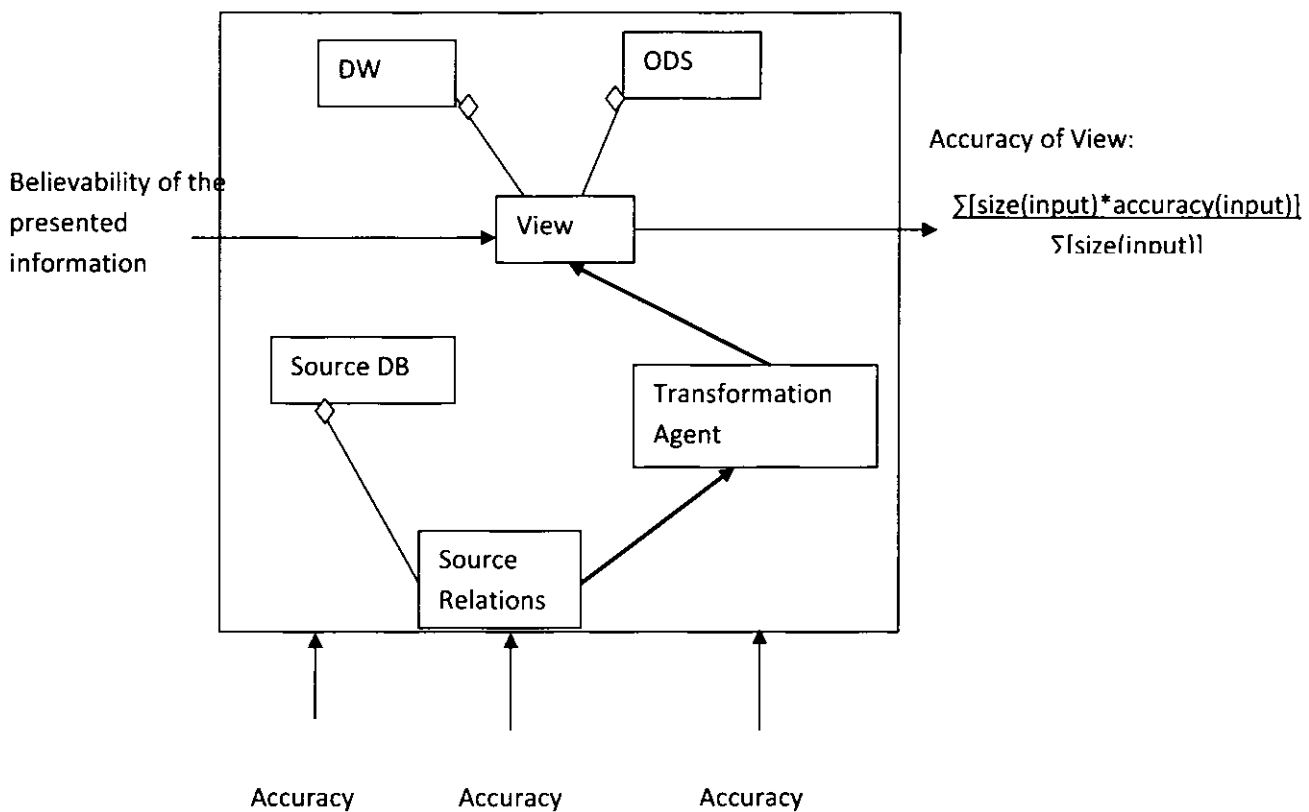
## 5.4 Case Study:

Let's take a simple 'Accuracy' case, suppose the requirement is 'the believability of the information delivered to the final user'.

To achieve the goal believability, we further specified the quality goal accuracy, the accuracy of the data in the views used by the final user. Accuracy is participating all the components in the refreshment of a view like DW, ODS, Agents (which convert the data to the desired format).

For the accuracy of the view we also provide an analytical function calculating it from the accuracy of the input data and from the size.

Quality goal scenario is presented in the figure below:



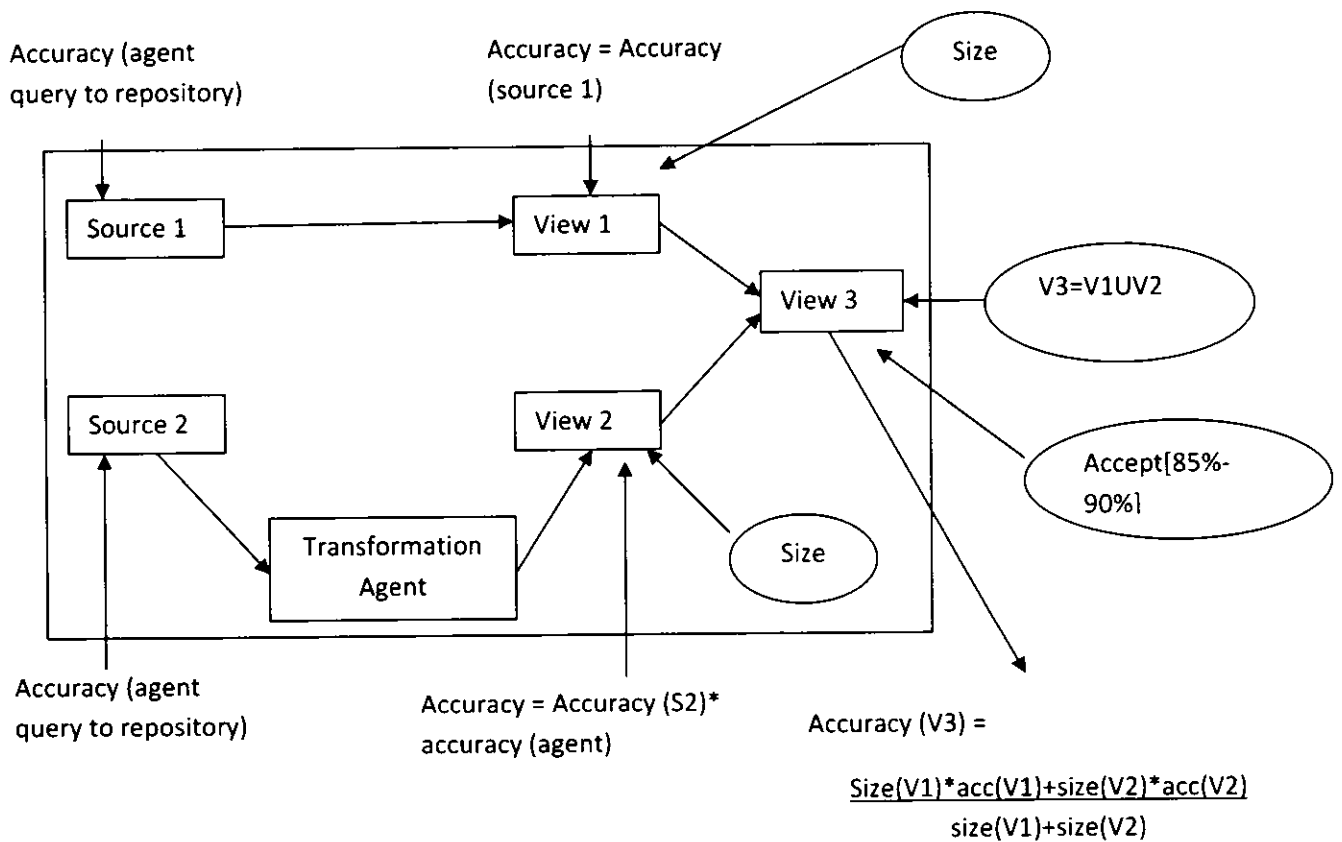
We can improve the believability of the view information by increasing or improving its level of accuracy through the improvement of the accuracy of the transformation agents or the source relations.

### 5.5 Evaluation of the Quality Goal:

After the Goal specification we must determine the specific object instances (relation, view) for the goal evaluation through a query to the Meta data repository.

In the above case (believability) suppose we identify two source relations (S1, S2) pumping data to two views (V1, V2), from these two views we derived another final view V3, the accuracy of which we have to quantify.

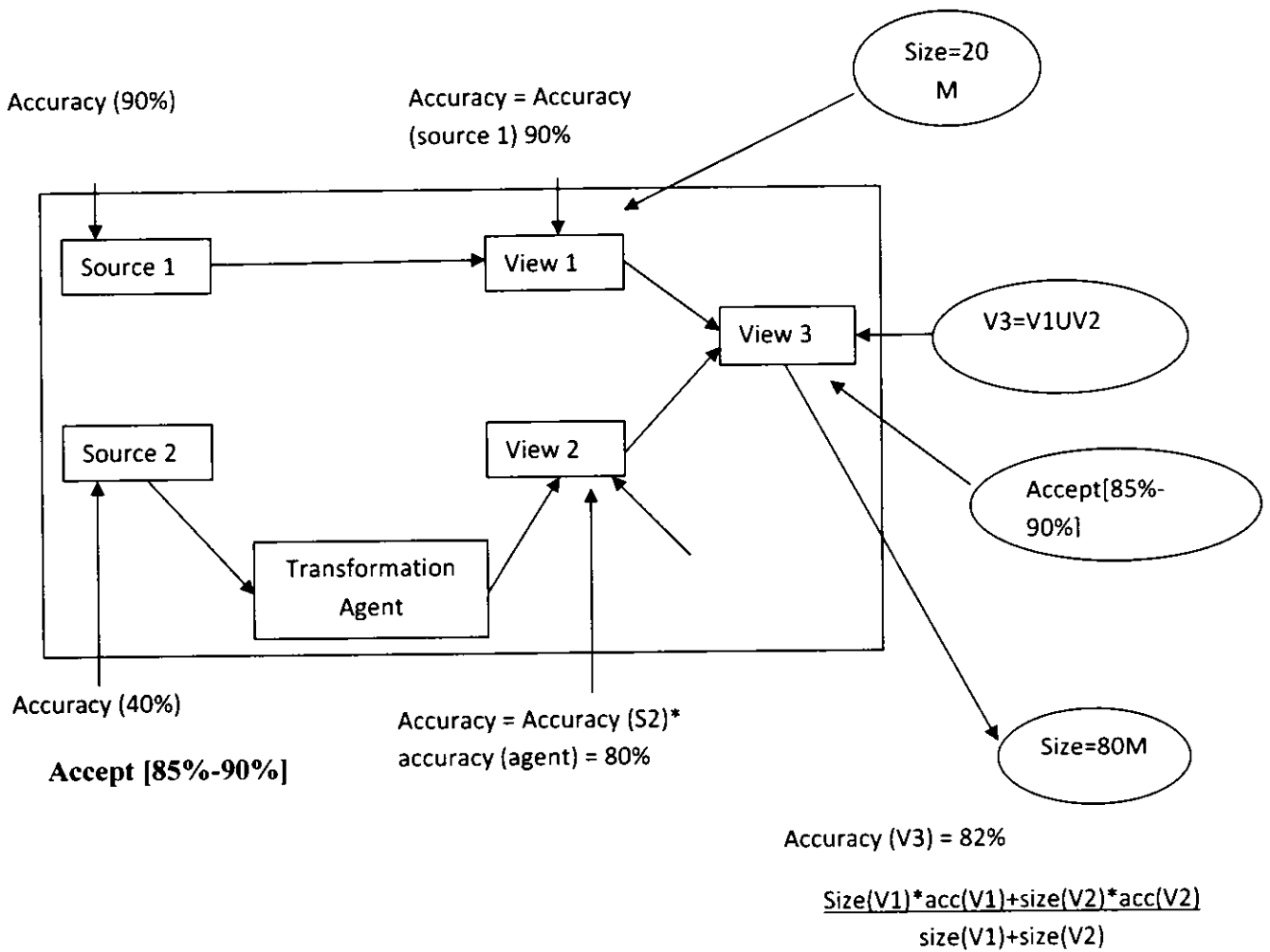
Next thing we must consider is several design choices i.e. the properties of the interplaying objects which greatly influence the quality of the result. In our case as we are dealing only with the size of the data so we can take into account the size of the propagated data and the view definitions at the regularity of the refreshment.



One should also determine the measuring agents for quality factors, if there's no agent defined then some computation procedure must be determine for calculating the actual values of the quality factors. Also, the parameters of the measuring procedures should be set accordingly.

The final step is to add the expected or acceptable values wherever necessary. For objective judgment of subjective quality goal, accepted range of values is the criteria whose result gives the well-defined map of problem to the stakeholder.

Then on the specified quality factor calculation or certain computation of the values are calculated. These values are obtained through the already defined agent. These values may be already recorded in the meta-data repository or can be obtained through the query to the meta-data repository.



Here, we considered meta-data repository is regularly refreshed through an external agent, by this several steps are omitted for simplicity of the case.

## 5.6 Improvement Phase:

As View 3 accuracy is not in desired range that is acceptable, then quality manager decides to take an action on such undesired situation.

The current GDI model can't express the large number of quality factors of organization. The consequence is that there is no systematic understanding of the interplay between quality factors and design options in data warehousing.

The worth of our approach is that it enables us to understand the mechanism which produces the problem. By using the quality functions we can easily figure out the problematic areas.

For example;

From the above scenario we can suggest following action;

a) Increase the 'view 1' and 'view 2' accuracy by 10% which implies that:

- increase accuracy of source 1 by 10%
- increase the accuracy of source 2 by 5 or 10 %

We can also determine the size of input views through the use of the specific measurements.

## 5.7 Summary:

In this chapter, implementation support of the enhanced model is proposed, in which different goals are selected and tested through a mathematical expression. A case study has done with the quality goal 'Accuracy'. Goal has evaluated and results have shown.

The selected goal 'Accuracy' has tested on which decision has applied that what kind of correction factors should be selected to change or effect the selected quality goal. In the last sixth chapter comparison between old and new approaches is given. Also conclusion and future work has been discussed in it



## **CHAPTER NO 6**

### **CONCLUSION & FUTURE WORK**

## 6.1 Comparison between Old & New Approaches Of Enhanced Goal Decision Information Approach

### Requirements of Data Warehouse:

Data warehouse system should have reliability, consistency and usefulness

<b>RELIABILITY</b>	<ul style="list-style-type: none"> <li>• <b>FRESHNESS</b></li> <li>• <b>ACCURACY</b></li> </ul>
<b>CONSISTENCY</b>	<ul style="list-style-type: none"> <li>• <b>CONSISTENCY</b></li> </ul>
<b>USEFULNESS</b>	<ul style="list-style-type: none"> <li>• <b>ACCESSIBILITY</b></li> <li>• <b>TIMELINESS</b></li> </ul>

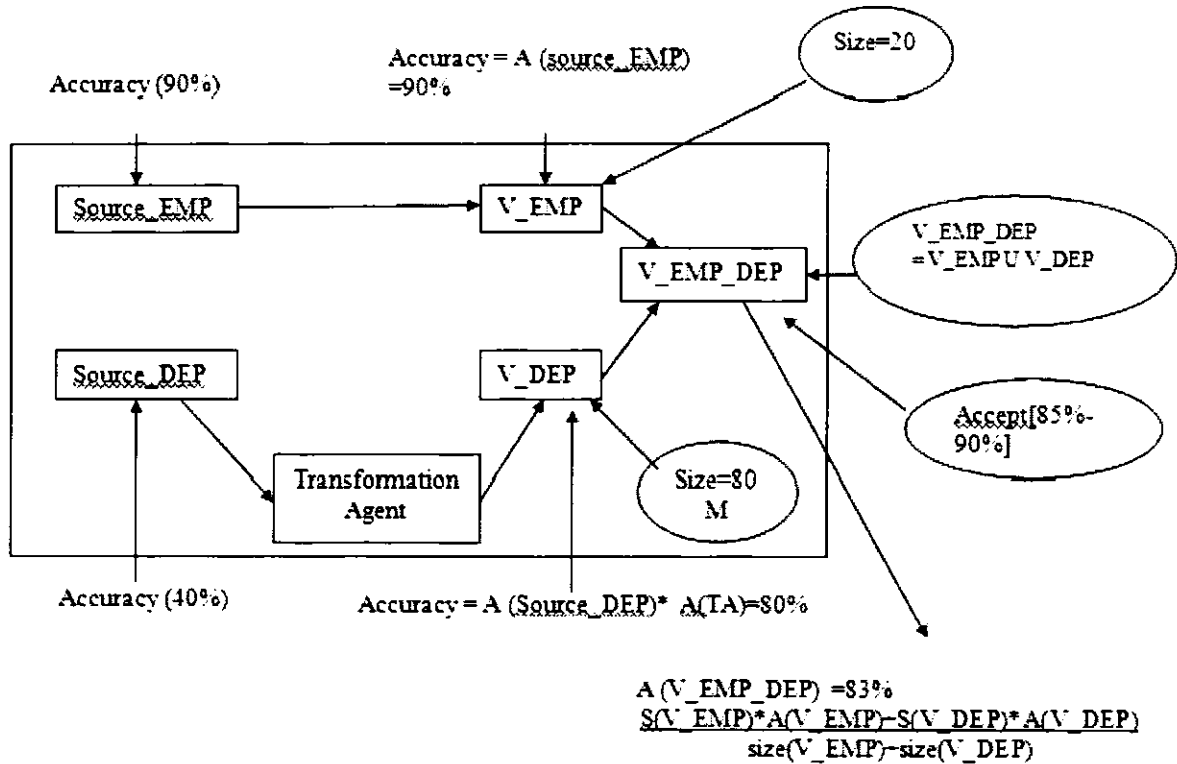
### Meta data of Data Sources Employee & Department (Random Value)

<b>Data Sources</b>	<b>Freshness</b>	<b>Accuracy</b>	<b>Consistency</b>	<b>Accessibility</b>	<b>Timeliness</b>
Employee	70%	90%	70%	80%	75%
Department	80%	40%	70%	70%	80%

“Check Reliability, consistency, usefulness of Source Employee & Department.”

Range [70 % – 90%]

**Old Model is unable to operate Goals Reliability & Usefulness due to their Complexity**



**2<sup>nd</sup> Check**

“Check Completeness, Availability, Correctness, Accuracy”

“Goals (Completeness, availability, correctness) can't be evaluated as they're not in business rules”. As they're Conflicting attribute and are irrelevant with our organizational business trends.

Accuracy                      10/cycle

**Therefore, 10CMC**

Goal Evaluation through Old Model

“Check Completeness, Availability, Correctness, Accuracy”.

**Evaluation Done as per below results**

Completeness	10/cycle
Availability	10/cycle
Correctness	10/cycle
Accuracy	<u>10/cycle</u>
Total	<b>50 CMC</b>

**Therefore, 50CMC**

So from above mentioned checks we concluded that old model is testing irrelevant goals which are not pre defined by our quality manager and also the computational cost (50 CM) is very high due to evaluation of conflicting quality attributes.

## 6.2 Conclusion & Future Work:

People usually think that Quality of data warehouse term is related to the quality of data, however quality with respect to data is very important but in fact the term quality refers to a bigger picture as stated below by David and Thomann.

According to David Wells and James Thomann;

“Quality is the act of measuring the progress of the data warehouse in terms of its ability to satisfy your customer base”.

Data quality is not enough- decisions are made on information quality and not only on data quality and data is only fit to be treated as a referenced for data quality measurements.

Here we deal with the issue of quality-oriented design usage and evolution of data warehouses. We have followed the approach of previous work [13], semantically rich meta-information of a data warehouse is stored in meta-data repository concerning the conceptual, logical and physical perspective of the data warehouse. In addition, the information on the quality of the stored objects is recorded in this repository.

Our approach extends GDI, based on the idea that a goal is operationally defined over a set of questions and requirements of the organization on the quality factors: this way the meta-data repository of data warehouse is not simply defined statically but is actually exploited in a systematic manner, which aims to map a high level subjective quality goal into the measurement of a set of interrelated quality factors.

The benefit from the use of the methodology is not only the obtained solution to a specific problem, may be of greater importance is the fact that the involved stakeholder gets a more clear view of the data warehouse interdependencies. This is achieved through the systematic methodological steps which convert a subjective problem to specific measurable quality factors that affect the solution to the problem and by which we may be able to measure the magnitude of the problem.

## **REFERENCES & BIBLIOGRAPHY**

## References

1. Data Warehouse Quality management (coursebook DWH-DWQ-004), by Laura Hadley 2007.
2. Matthias Jarke, Manfred A. Jeusfeld, Christoph Quix, Panos Vassiliadis., Architecture and quality in data warehouses: An Extended Repository Approach. 52056 Aachen, Germany 2006.
3. Becker, D., McMullen, W., Hetherington-Young, K. 2007. A Flexible and Generic Data Quality Metamodel. In Proc. International Conference on Information Quality 2007 (MIT, Cambridge, MA, USA, 2007). ICIQ 2007. Internet: <http://mitiq.mit.edu/iciq/ICIQ/iqpapers.aspx?iciqyear=2007>.
4. Farinha, J., and Trigueiros, M. J. 2007. An Extensible Metadata Framework for Data Quality Assessment of Composite Structures. In Proc. Data Warehouse and Knowledge (Regensburg, Germany, 2007). LNCS, vol. 4654, 34-44. DaWaK'07. Springer, Berlin/Heidelberg, Germany.
5. Gomes, P., Farinha, J., Trigueiros, M.J.: A Data Quality Metamodel Extension to CWM. In Proc. 4th Asia-Pacific Conference on Conceptual Modelling (Ballarat, Australia 2007). APCCM 2007. CRPIT, 67. ACS, Darlinghurst, Australia, 17-26.
6. Prakash N., Singh Y., Gosain A. (2005): Requirement Driven Approach for Development of Banking Data Warehouse. Journal of the CSI on 35 (3).
7. Ranjit Singh, D. K. S. (May 2010). "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing." IJCSI International Journal of Computer Science Issues Vol. 7(Issue 3).
8. José Farinha, M. J. T., Orlando Belo (2009). "Using Inheritance in a Metadata Based Approach to Data Quality Assessment." MoSE+DQS'09, November 6, 2009, Hong Kong, China. Copyright 2009 ACM 978-1-60558-816-2/09/11 (ISBN: 978-1-60558-816-2).
9. Anjana Gosain, J. S. U. S. o. I. T. and D. Guru Gobind Singh Indrarastha University, India (April 2009). "Achieving Data Warehouse Quality Using GDI Approach." International Journal of Information Studies Volume 1(Issue 2).
10. Amer Nizar AbuAli, H. Y. A.-A. (2010). "Data Warehouse Critical Success Factors." European Journal of Scientific Research vol. 42 (ISSN 1450-216X Vol.42 No.2 (2010), pp.326-335).
11. Data Warehouse Evolution Framework, <http://syrcondis.citforum.ru/2007/13.pdf>. Accessed Data: Oct 11, 2008

12. Dittrich, A. V. K. R. (2001). "Metadata Management for Data Warehousing: Between Vision and Reality. ." ideas, pp.0129, 2001 International Database Engineering & Applications Symposium (IDEAS '01)
13. Nazih Selmoune Zaia Alimazighi "A decisional tool for quality improvement in Higher Education", 3rd International Conference on Information & Communication Technologies, Vol.4.
14. 14. Timon C. Du, Jacqueline Wong , "Designing Data Warehouses for Supply Chain Management" , 2004 Proceedings of the IEEE International Conference on E-Commerce Technology.
15. Karen C. Davis, Il-Yeol Song, "Data Warehousing and OLAP", Journal of Database Management, Vol.17, No.1, 2006, pp.1-3.
16. Jane Zhao , "Designing Distributed Data Warehouses and OLAP Systems" , Journal of Systems and Software May 2005, Volume 79 , Issue 5
17. Panos Vassiliadis<sup>1</sup>, Mokrane Bouzeghoub<sup>2</sup>, Christoph Quix<sup>3</sup>. "Towards Quality-Oriented Data Warehouse Usage and Evolution" Information Systems Volume 25, Issue 2, April 2000, Pages 89-115 The 11th International Conference on Advanced Information System Engineering .