Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-8, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

Data Quality Problems in Data Warehousing With Descriptive Classification

D. KANMANI[a] & G. EZHILARASI [b]

[a] Research Scholar, Department of Computer Science, PRIST University, Thanjavur.

[b] Research Supervisor, Department of Computer Science, PRIST University, Thanjavur.

***Abstract**: Data warehousing is the process of constructing of data. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. There is one key obstacle to the rapid development and implementation of quality data warehouses specifically that of warehouse data quality issues at various stages of data warehousing. Specifically, problems arise in populating a warehouse with quality data. The problems arise in populating a warehouse with quality data.Morever the period of time many researchers have contributed to the data quality issues, but yet we didn't identify the causes of data quality problems at all the phases of data warehousing Viz. 1) data sources, 2) data integration & data profiling, 3) Data staging and ETL, 4) data warehouse modelling & schema design. The purpose of the paper is to identify the reasons for data deficiencies, non-availability or reach ability problems at all the above mentioned stages of data warehousing and to give some classification of these causes as well as solution for improving data quality through Statistical Process Control (SPC), Quality engineering management.*

***Keywords**- ETL, SPC, heterogeneous*

I. INTRODUCTION

Data Warehouse plays an important part in the process of knowledge engineering and decision-making for Enterprise, as a key component of the data warehouse architecture, the tool that support data extraction, transformation, loading (ETL) is a critical success factor for any data warehouse projects Traditional methods of ETL development The existence of data alone does not ensure that all the management functions and decisions can be smoothly undertaken. A broader definition is that data quality is achieved when organization uses data that is comprehensive, understandable, consistent, relevant and timely. Understanding the key data quality dimensions is the first step to data quality improvement. Abundant attempts have been made to define data quality and to identify its dimensions. Dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness; understand ability, conciseness and usefulness. For our research paper we have under taken the quality criteria by taking 6 key dimensions as depicted below figure1.

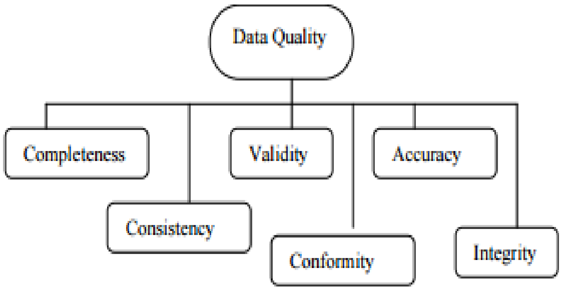


Figure 1: Data Quality Criteria [21]

1.1 ENSURING DATA QUALITY

Data quality is an increasingly serious issue for organizations large and small. It is central to all data integration initiatives. Before data can be used effectively in a data warehouse, or in customer relationship management, enterprise resource planning or business analytics applications, It need to be analyzed and cleansed .Understanding the key data quality dimensions is the first step to data quality improvement. To be process able and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. Abundant attempts have been made to define data quality and to identify its dimensions. Dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness; understand ability, conciseness and usefulness. For this research paper I have under taken the quality criteria by taking 6 key dimensions as • Completeness • Consistency • Validity •

Conformity • Accuracy • Integrity • Completeness: deals with to ensure is all the requisite information available? Are some data values missing, or in an unusable state? Consistency: Do distinct occurrences of the same data instances agree with each other or provide conflicting information. Are values consistent across data sets? Validity: refers to the correctness and reasonableness of data Conformity: Are there expectations that data values conform to specified formats? If so, do all the values Accuracy: Do data objects accurately represent the "real world" values they are expected to model? Incorrect spellings of product or person names, addresses, and even untimely or not current data can impact operational and analytical applications. Integrity: What data is missing important relationship linkages? The inability to link related records together may actually introduce duplication

Data warehouses are one of the foundations of the Decision Support Systems of many IS operations. As defined by the "father of data warehouse", William H. Inman, a data warehouse is "a collection of Integrated, Subject-Oriented, Non Volatile and Time Variant databases where each unit of data is specific to some period of time. Data Warehouses can contain detailed data, lightly summarized data and highly summarized data, all formatted for analysis and decision support" (Inman, 1996). In the "Data Warehouse Toolkit", Ralph Kimball gives a more concise definition: "a copy of transaction data specifically structured for query and analysis" (Kimball, 1998). Both definitions stress the data warehouse's analysis focus, and highlight the historical nature of the data found in a data warehouse.

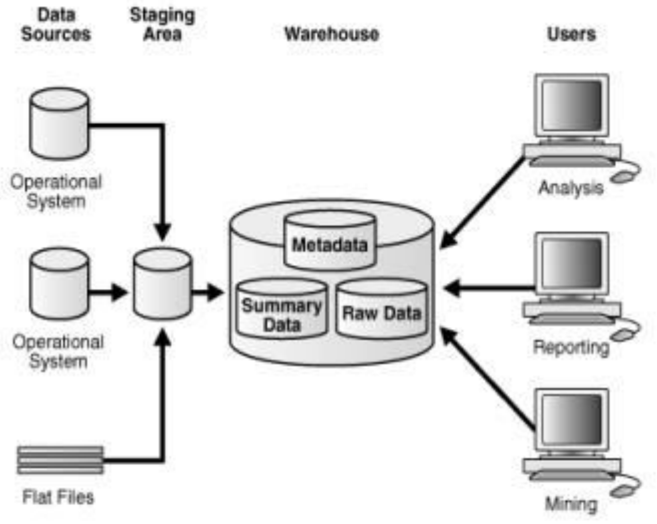


Figure 2: Data Warehousing Structure

II. DATA WAREHOUSING AND THE MAIN MODULE OF ETL

Data warehouses are one of the foundations of the Decision Support Systems of many IS operations. Data Warehouses can contain detailed data, lightly summarized data and highly summarized data, all formatted for analysis and decision support".Accurate way to design ETL process as in fig. to make it efficient, flexible and maintainable, ETL can be divided into 5 modules: data extraction, data validation, data cleaning, data conversion and data loading. 1] Data extraction This step requires a great deal, first of all, need to find out which business system does the business data come from, what kind of database management system the business database server runs. Secondly, need to know what kind of table structure the database has and the corresponding meaning of each table structure. Thirdly, need to check whether there exist the manual data and the quantity of the data. Fourthly, define whether there exist the unstructured data. After collecting all of the information, give out he data explanation file. 2] Data validation Data validation involves a lot of checking work, including the effective value of the property, foreign key checking and so on. As for the low-quantity data, refuse them firstly, and then these data will be stored in order to be fixed in the field of the amendment. 3] Data cleaning The task of data cleaning is to filter out the undesirable data, and then send them to the business operation department. These undesirable data include: incomplete data, wrong data, duplicated data and so on. 4] Data conversion From the micro-details perspective, data conversion involves the following types: direct mapping, field operations, character string processing, null value determination, date conversion, date operation, and assemble operations and so on. 5] Data loading Data will be moved to the center of the target datawarehouse table, it is usually the last step in the process of ETL. As to the best way to load data, the implementation depends on the type of operation and the quantity of data. There are two ways to insert or update data in the database table: SQL insert/update/delete or batch loading application program.

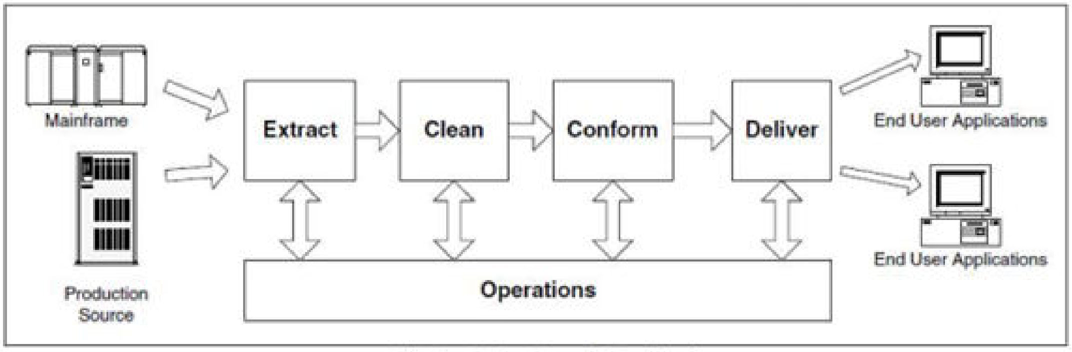


Fig .3 .The Four Staging Steps of a Data Warehouse.

III. LITERATURE REVIEW

ETL is the core process of building data warehouse and is developed by and based on operation system. The data stored in data warehouse is from transaction system and external data sources.

The study is designed as a literature review of materials published between 1992 and 2008 on the topics of data quality and data warehouses. The Figure presents the resulting research model formulated trough extensive literature review. To develop the research model, the IT implementations infrastructure, data warehousing literature, research questionnaires related to data quality were reviewed to identify various reasons of data duality problems at the stages mentioned in the model. Classification of causes of data quality problems so formed will be divided into the factors responsible for data quality at the phases. Later in the next phase of study, it will be converted into survey instrument for the confirmation of these issues from the data warehouse practitioners.

1.Won Kim et al (2002) paper presented a comprehensive taxonomy of dirty data and explored the impact of dirty data on data mining results. A comprehensive classification of dirty data developed for use as a framework for understanding how dirty data arise, manifest themselves, and may be cleansed to ensure proper construction of data warehouses and accurate data analysis.

2. AmitRudra and Emilie Yeo (1999) the paper concluded that the quality of data in a data warehouse could be influenced by factors like: data not fully captured, heterogeneous system integration and lack of policy and planning from management.

3. Channah E Naiman&Aris M. Ouksel (1995)- the paper proposed a classification of semantic conflicts and highlighted the issue of semantic heterogeneity, schema integration problems which further may have far reaching consequences on data quality . John Hess (1998) the report has highlighted the importance of handling of missing values of the data sources, specially emphasized on missing dimension attribute values.

4. Scott W. Ambler (2001) the article explored the wide variety of problems with the legacy data, including data quality, data design, data architecture, and political/process related issues. The article has provided a brief bifurcation of common issues of legacy data, which contribute to the data quality problems

5. JaideepSrivastava, Ping-Yao Chen (1999) the principal goal of this paper is to identify the common issues in data integration and data-warehouse creation. Problems arise in populating a warehouse with existing data since it has various types of heterogeneity.

6. Wayne Eckerson (2004) data warehousing projects gloss over the all-important step of scrutinizing source data before designing data models and ETL mappings. The paper presented

the reasons for data quality problems out of which most important are 1) Discovering Errors Too Late 2) Unreliable Meta Data. 3) Manual Profiling. 4) Lack of selection of automated profiling tools.

3.1.STAGES OF DATA WAREHOUSING SUSCEPTIBLE TO DATA QUALITY PROBLEMS

The purpose is to formulate a descriptive taxonomy of all the issues at all the stages of Data Warehousing. The phases are: Data Source• Data Integration and Data Profiling• Data Staging and ETL•

3.1.1 Data Quality Issues at Data Sources

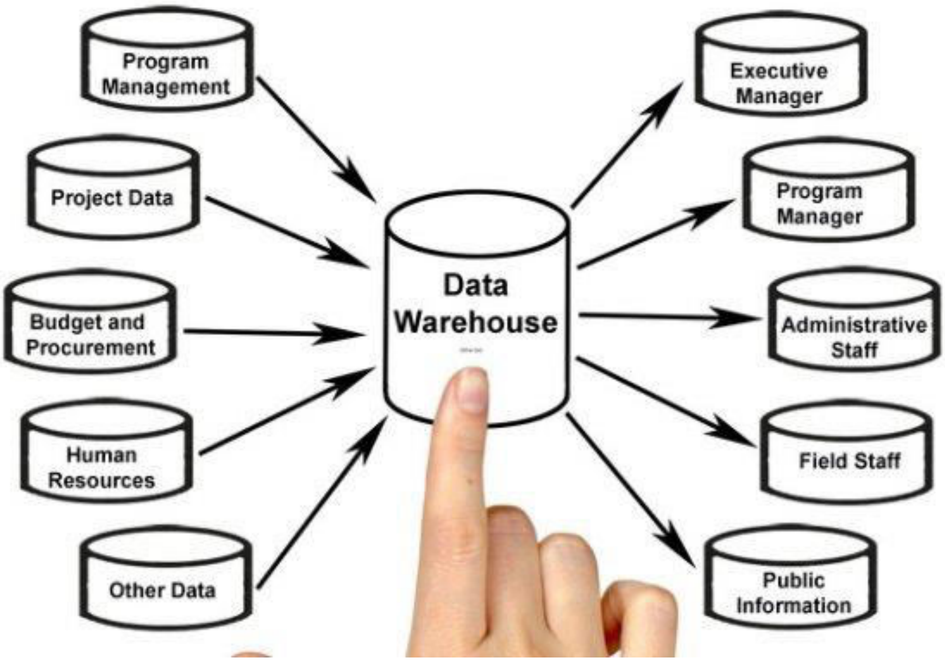
Different data Sources have different kind of problems associated with it such as data from legacy data sources (e.g., mainframe-based COBOL programs) do not even have metadata that describe them. The sources of dirty data include data entry error by a human or computer system, data update error by a human or computer system. Part of the data comes from text files, part from MS Excel files and some of the data is direct ODBC connection to the source database.

A leading cause of data warehousing and business intelligence project failures is to obtain the wrong or poor quality data. Eventually data in the data warehouse is fed from various sources as depicted .

Some of the data sources are cooperative and some might be non cooperative sources. Because of this diversity several reasons are present which may contribute to data quality problems, if not properly taken care of. A source that offers any kind of unsecured access can become unreliable-and ultimately contributing to poor data quality.

The source system consists of all those 'transaction/Production' raw data providers, from where the details are pulled out for making it suitable for Data Warehousing. All these data sources are having their own methods of storing data.

Fig:4:Data warehouse



Consider the following seven sources of data quality issues. 1. Entry quality: Did the information enter the system correctly at the origin? 2. Process quality: Was the integrity of the information maintained during processing through the system? 3. Identification quality: Are two similar objects identified correctly to be the same or different? 4. Integration quality: Is all the known information about an object integrated to the point of providing an accurate representation of the object? 5. Usage quality: Is the information used and interpreted correctly at the point of access? 6. Aging quality: Has enough time passed that the validity of the information can no longer be trusted? 7. Organizational quality: Can the same information be reconciled between two systems based on the way the organization constructs and views the data? A plan of action must account for each of these sources of error. Each case differs in its ease of detection and ease of correction. An examination of each of these sources reveals a varying amount of costs associated with each and inconsistent amounts of difficulty to address the problem.

Entry Quality Entry quality is probably the easiest problem to identify but is often the most difficult to correct. Entry issues are usually caused by a person entering data into a system. The problem may be a typo or a willful decision, such as providing a dummy phone number or address. Identifying these outliers or missing data is easily accomplished with profiling tools or simple queries. The cost of entry problems depends on the use. If a phone number or email address is used only for informational purposes, then the cost of its absence is probably low. If instead, a phone number is used for marketing and driving new sales, then opportunity cost may be significant over a major percentage of records. Addressing data quality at the source can be difficult. If data was sourced from a third party, there is usually little the organization can do. Likewise, applications that provide internal sources of data might be old and too expensive to modify. And there are few incentives for the clerks at the point of entry to obtain, verify, and enter every data point. Process Quality Process quality issues usually occur systematically as data is moved through an organization. They may result from a system crash, lost file, or any other technical occurrence that results from integrated systems. These issues are often difficult to identify, especially if the data has made a number of transformations on the way to its destination. Process quality can usually be remedied easily once the source of the problem is identified. Proper checks and quality control at each touch point along the path can help ensure that problems are rooted out, but these checks are often absent in legacy processes. Identification Quality Identification quality problems result from a failure to recognize the relationship between two objects. For example, two similar products with different SKUs are incorrectly judged to be the same. Identification quality may have significant associated costs, such as mailing the same household more than once. Data quality processes can largely eliminate this problem by matching records, identifying duplicates and placing a confidence score on the similarity of records. Ambiguously scored records can be reviewed and judged by a data steward. Still, the results are never perfect, and determining the proper business rules for matching can involve trial and error.

Integration Quality Integration quality, or quality of completeness, can present big challenges for large organizations. Integration quality problems occur because information is isolated by system or departmental boundaries. It might be important for an auto claims adjuster to know that a customer is also a high-value life insurance customer, but if the auto and life insurance systems are not integrated, that information will not be available.

Aging Quality The most challenging aspect of aging quality is determining at which point the information is no longer valid. Usually, such decisions are somewhat arbitrary and vary by usage. For example, maintaining a former customer's address for more than five years is probably not useful. If customers haven't been heard from in several years despite marketing efforts, how can we be certain they still live at the same address? At the same time, maintaining customer address information for a homeowner's insurance claim may be necessary and even required by law. Such decisions need to be made by the business owners and the rules should be architected into the solution.

Usage Quality Usage quality often presents itself when data warehouse developers lack access to legacy source documentation or subject matter experts. Without adequate guidance, they are left to guess the meaning and use of certain data elements. Another scenario occurs in organizations where users are given the tools to write their own queries or create their own reports. Incorrect usage may be difficult to detect and quantify in cost

3.1.2 A Strategic Approach

The first step to developing a data strategy is to identify where quality problems exist. After identifying the problem, it is important to assess the business impact and cost to the organization

In addition, the cost associated with a particular issue may be small at a departmental level but much greater when viewed across the entire enterprise.

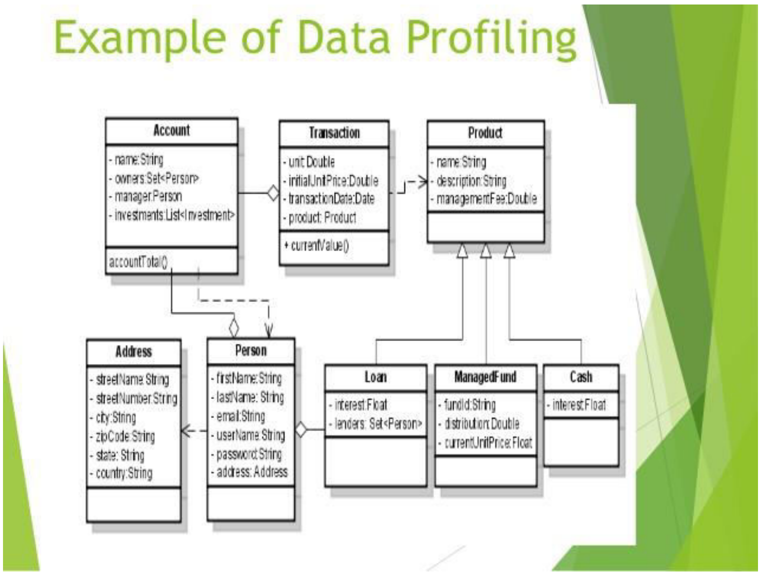
Addressing data quality requires changes in the way we conduct our business and in our technology framework. It requires organizational commitment and long-term vision.

Addressing data quality requires changes in the way we conduct our business and in our technology framework. It requires organizational commitment and long-term vision.

3.1.3 Causes of Data Quality Issues at Data Profiling

When possible candidate data sources are identified and finalized data profiling comes in play immediately. Data profiling is the examination and assessment of your source systems' data quality, integrity and consistency sometimes also called as source systems analysis.

Fig:5 :causes of data quality at Data profiling



3.1.4 ETL (Extract-Transform-Load)

ETL comes from Data Warehousing and stands for Extract-Transform-Load. ETL covers a process of how the data are loaded from the source system to the data warehouse. Currently, the ETL encompasses a cleaning step as a separate step. The sequence is then Extract-Clean-Transform-Load. Let us briefly describe each step of the ETL process.

3.2 PROCESS

3.2.1 Extract

The Extract step covers the data extraction from the source system and makes it accessible for further processing. The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible. The extract step should be designed in a way that it does not negatively affect the source system in terms or performance, response time or any kind of locking.

There are several ways to perform the extract:

- Update notification - if the source system is able to provide a notification that a record has been changed and describe the change, this is the easiest way to get the data.
- Incremental extract - some systems may not be able to provide notification that an update has occurred, but they are able to identify which records have been modified and provide an extract of such records. During further ETL steps, the system needs to identify changes and propagate it down. Note, that by using daily extract, we may not be able to handle deleted records properly.
- Full extract - some systems are not able to identify which data has been changed at all, so a full extract is the only way one can get the data out of the system. The full extract requires keeping a copy of the last extract in the same format in order to be able to identify changes. Full extract handles deletions as well.

When using Incremental or Full extracts, the extract frequency is extremely important. Particularly for full extracts; the data volumes can be in tens of gigabytes.

3.2.2 Clean

The cleaning step is one of the most important as it ensures the quality of the data in the data warehouse. Cleaning should perform basic data unification rules, such as:

- Making identifiers unique (sex categories Male/Female/Unknown, M/F/null, Man/Woman/Not Available are translated to standard Male/Female/Unknown)
- Convert null values into standardized Not Available/Not Provided value
- Convert phone numbers, ZIP codes to a standardized form
- Validate address fields, convert them into proper naming, e.g. Street/St/St./Str./Str
- Validate address fields against each other (State/Country, City/State, City/ZIP code, City/Street).

3.2.3 Transform

The transform step applies a set of rules to transform the data from the source to the target. This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined. The transformation step also requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

3.2.4 Load

During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database. In order to make the load process efficient, it is helpful to disable any constraints and indexes before the load and enable them back only after the load completes. The referential integrity needs to be maintained by ETL tool to ensure consistency.

3.3 MANAGING ETL PROCESS

The ETL process seems quite straight forward. As with every application, there is a possibility that the ETL process fails. This can be caused by missing extracts from one of the systems, missing values in one of the reference tables, or simply a connection or power outage. Therefore, it is necessary to design the ETL process keeping fail-recovery in mind.

3.3.1 Staging

It should be possible to restart, at least, some of the phases independently from the others. For example, if the transformation step fails, it should not be necessary to restart the Extract step. We can ensure this by implementing proper staging. Staging means that the data is simply dumped to the location (called the Staging Area) so that it can then be read by the next processing phase. The staging area is also used during ETL process to store intermediate results of processing. This is ok for the ETL process which uses for this purpose. However, The staging area should is be accessed by the load ETL process only. It should never be available to anyone else; particularly not to end users as it is not intended for data presentation to the end-user.may contain incomplete or in-the-middle-of-the-processing data.

3.4 ETL TOOL IMPLEMENTATION

When you are about to use an ETL tool, there is a fundamental decision to be made: will the company build its own data transformation tool or will it use an existing tool?

Building your own data transformation tool (usually a set of shell scripts) is the preferred approach for a small number of data sources which reside in storage of the same type. The reason for that is the effort to implement the necessary transformation is little due to similar data structure and common system architecture. Also, this approach saves licensing cost and there is no need to train the staff in a new tool. This approach, however, is dangerous from the TOC point of view. If the transformations become more sophisticated during the time or there is a need to integrate other systems, the complexity of such an ETL system grows but the manageability drops significantly. Similarly, the implementation of your own tool often resembles re-inventing the wheel.

There are many ready-to-use ETL tools on the market. The main benefit of using off-the-shelf ETL tools is the fact that they are optimized for the ETL process by providing connectors to common data sources like databases, flat files, mainframe systems, xml, etc. They provide a means to implement data transformations easily and consistently across various data sources. This includes filtering, reformatting, sorting, joining, merging, aggregation and other operations ready to use. The tools also support transformation scheduling, version control, monitoring and unified metadata management. Some of the ETL tools are even integrated with BI tools.

IV .Conclusion

In this paper attempt has been made to collect all possible causes of data quality problems that may exist at all the phases of data warehouse. My objective was to put forth such a descriptive classification which covers all the phases of data warehousing which can impact the data quality.

And also to provide solution for improving Data quality. The motivation of the research was to integrate all the sayings of different researches which were focused on individual phases of data warehouse. Such as lot of literature is available on dirty data taxonomies.

REFERENCES:

[1] Channah F. Naiman, Aris M. Ouksel (1995) "A Classification of Semantic Conflicts in Heterogeneous Database Systems", Journal of Organizational Computing, Vol. 5, 1995 [2] John Hess (1998), "Dealing With Missing Values In The Data Warehouse" A Report of Stonebridge Technologies, Inc (1998). [3] Jaideep Srivastava, Ping-Yao Chen (1999) "Warehouse Creation-A Potential Roadblock to Data Warehousing", IEEE Transactions on Knowledge and Data Engineering January/February 1999 (Vol. 11, No. 1) pp. 118-126

[2].Simitsis, A.; Wilkinson, K.; Dayal, U.; Castellanos, M.; "Optimizing ETL workflows for fault-tolerance",Data Engineering (ICDE), 2010 IEEE 26th International Conference on Digital Object Identifier: 10.1109/ICDE.2010.5447816 Publication Year: 2010 ,Page(s): 385 – 396

[3]. Wayne W. E. (2004) "Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data ", The Data warehouse Institute (TDWI) report, available at www.dw-institute.com .