IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing

Ranjit Singh[1], Dr. Kawaljeet Singh[2]

[1] Research Scholar, University College of Engineering (UCoE), Punjabi University
Patiala (Punjab), INDIA

[2] Director, University Computer Center (UCC), Punjabi University
Patiala (Punjab), INDIA

Abstract

Data warehousing is gaining in eminence as organizations become awake of the benefits of decision oriented and business intelligence oriented data bases. However, there is one key stumbling block 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. Over the period of time many researchers have contributed to the data quality issues, but no research has collectively gathered all 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 modeling & schema design. The state-of-the-art purpose of the paper is to identify the reasons for data deficiencies, non-availability or reach ability problems at all the aforementioned stages of data warehousing and to formulate descriptive classification of these causes. We have identified possible set of causes of data quality issues from the extensive literature review and with consultation of the data warehouse practitioners working in renowned IT giants on India. We hope this will help developers & Implementers of warehouse to examine and analyze these issues before moving ahead for data integration and data warehouse solutions for quality decision oriented and business intelligence oriented applications.

Keywords : Data Quality (DQ), ETL, Data Staging, Data Warehouse

Section I – Introduction

1.1 Understanding Data Quality

The existence of data alone does not ensure that all the management functions and decisions can be smoothly undertaken. The one definition of data quality is that it's about bad data - data that is missing or incorrect or invalid in some context. 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. 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, understandability, conciseness and usefulness. For our research paper we have under taken the quality criteria by taking 6 key dimensions as depicted below figure1.

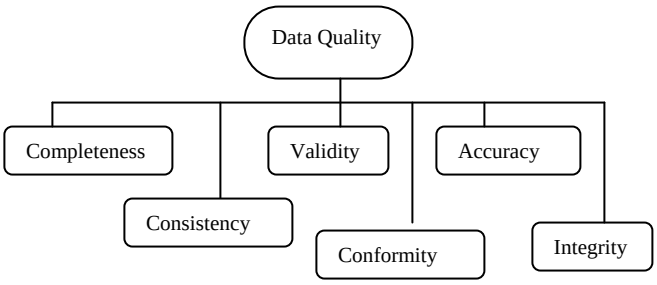


Figure 1: Data Quality Criteria [21]

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

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

conform to those formats? Maintaining conformance to specific formats is important.

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 across your systems.

1.2 Data Warehousing

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. Inmon, 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" (Inmon, 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.

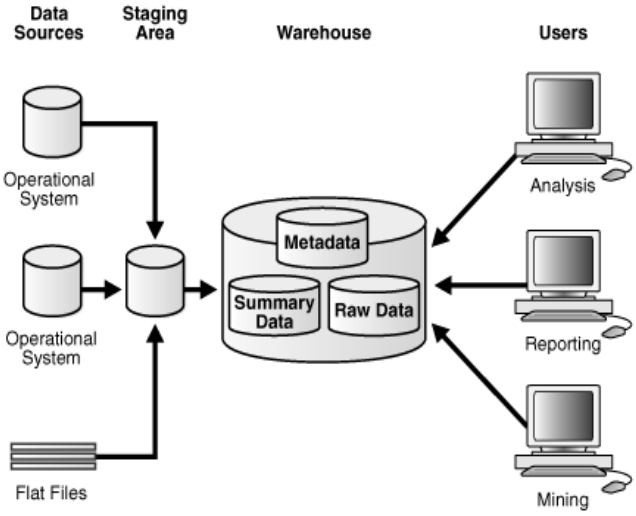


Figure 2: Data Warehousing Structure

1.3 Stages of Data Warehousing Susceptible to Data Quality Problems

The purpose of paper here 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
- Database Scheme (Modeling)

Quality of data can be compromised depending upon how data is received, entered, integrated, maintained, processed (Extracted, Transformed and Cleansed) and loaded. Data is impacted by numerous processes that bring data into your data environment, most of which affect its quality to some extent. All these phases of data warehousing are responsible for data quality in the data warehouse. Despite all the efforts, there still exists a certain percentage of dirty data. This residual dirty data should be reported, stating the reasons for the failure in data cleansing for the same.

Data quality problems can occur in many different ways [9]. The most common include:

- Poor data handling procedures and processes.
- Failure to stick on to data entry and maintenance procedures.
- Errors in the migration process from one system to another.
- External and third-party data that may not fit with your company data standards or may otherwise be of unconvinced quality.

The assumptions undertaken are that data quality issues can arise at any stage of data warehousing viz. in data sources, in data integration & profiling, in data staging, in ETL and database modeling. Following model is depicting the possible stages which are vulnerable of getting data quality problems

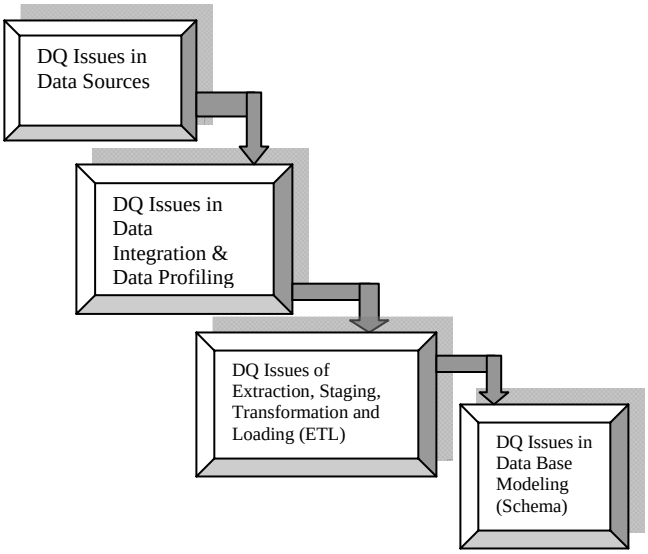


Figure 3: Stages of Data Warehouse Susceptible for DQ Problems

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

Section II

2.1 Methodology

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 Figure3 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.

2.1.1 Literature Reviewed.

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.

Jaideep Srivastava, 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.

Amit Rudra 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.

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.

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 is 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.

Wane Eckerson & Colin White (2003) report says ETL tools are the traffic cop for business intelligence applications. They control the flow of data between myriad source systems and BI applications. As BI environments expand and grow more complex, ETL tools need to change to keep pace. ETL tools need to evolve from batch-oriented, single-threaded processing that extracts and loads data in bulk to continuous, parallel processes that capture data in near real time. They also need to provide enhanced administration and ensure reliable operations and high availability.

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

2.2 Classification of Data Quality Issues

In order for the analyst to determine the scope of the underlying root causes of data quality issues and to plan the design the tools which can be used to address data quality issues, it is valuable to understand these common data quality issues. For the purpose of it the classification formed will be highly helpful to the data warehouse and data quality community.

2.2.1 Data Quality Issues at Data Sources

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 in the figure 4. 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. 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.

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,

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

part from MS Excel files and some of the data is direct ODBC connection to the source database [16].

Some files are result of manual consolidation of multiple files as a result of which data quality might be compromised at any step. Table 1 summarizes the possible causes of data quality problems at data sources stage of data warehousing.

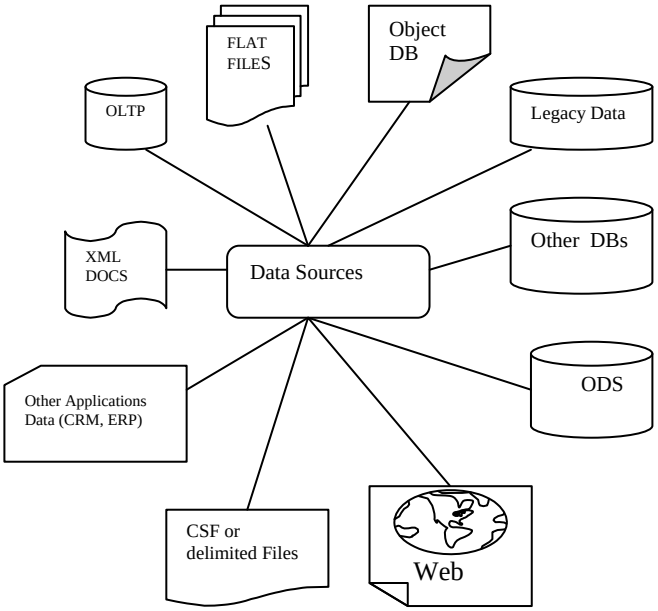


Fig 4: Possible Data Sources for Data Warehousing

Table 1:- Causes of Data Quality Problems at Data Sources Stage

Sr. No	*CAUSES OF DATA QUALITY PROBLEMS AT DATA SOURCES*
1	*Inadequate selection of candidate data sources cause DQ Problems (sources which do not comply to business rules)*
2	*As time and proximity from the source increase, the chances for getting correct data decrease [4].*
3	*In adequate knowledge of inter dependencies among data sources incorporate DQ problems.*
4	*Inability to cope with ageing data contribute to data quality problems.[4]*
5	*Varying timeliness of data sources [6] [7].*
6	*Lack of validation routines at sources causes DQ Problems.*
7	*Unexpected changes in source systems cause DQ Problems.*
8	*Multiple data sources generate semantic heterogeneity which leads to data quality issues [1][4]*
9	*The complexity of a data warehouse increases geometrically with the span of time of data to be fed into it*
10	*Usage of decontrolled applications and databases as data sources for data warehouse in the organizations.*
11	*Use of different representation formats in data sources.*
12	*Measurement errors [11].*
13	*Non-Compliance of data in data sources with the Standards.*
14	*Failure to update the sources in a timely manner causes DQ Problems.*
15	*Failure to update all replicas of data causes DQ Problems.*
16	*Presence of duplicate records of same data in multiple sources cause DQ Problems [6] [7] [11].*
17	*Approximations or surrogates used in data sources.*
18	*Contradictory information present in data sources cause data quality problems [6] [7].*
19	*Different encoding formats (ASCII, EBCDIC,….) [11]*
20	*Inadequate data quality testing on individual data source lead to poor data quality.*
21	*Lack of business ownership, policy and planning of the entire enterprise data contribute to data quality problems. [4]*
22	*Columns .having incorrect data values (for Eg. The AgeInYear Column for a person contain value 3 although the birthdate column is having value Aug, 14, 1967) [6] [7].*
23	*Having Inconsistent/Incorrect data formatting (The name of a person is stored in one table in the format "Firstname Surname" and in another table in the forma t "Surname, Firstname") [6] [7] [11].*
24	*System fields designed to allow free forms (Field not having adequate length).*
25	*Missing Columns (You need a middle name of a person but a column for it does not exist.) [6][7].*
26	*Missing values in data sources [2][11][12].*
27	*Misspelled data [11][12]*
28	*Additional columns [6] [7] [11].*
29	*Multiple sources for the same data ((For Eg. customer information is stored in three separate legacy databases)[11].*
30	*Various key strategies for the same type of entity (for Eg. One table stores customer information using the Social Security Number as the key, another uses the ClientID as the key, and another uses a surrogate key) [6] [7].*
31	*Inconsistent use of special characters (for Eg. A date uses hyphens to separate the year, month, and day whereas a numerical value stored as a string*

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

	uses hyphens to indicate negative numbers)[11] [20].
32	*Different data types for similar columns (A customer ID is stored as a number in one table and a string in another).*
33	*Varying default values used for missing data [6] [7].*
34	*Various representations of data in source data (The day of the week is stored as T, Tues, 2, and Tuesday in four separate Columns) [6] [7] [20].*
35	*Lack of record level validation in source Data.*
36	*Data values stray from their field description and business rules (Such as the Maiden Name column is being used to store a person's Hobbies, zip code into phone number box) [6] [7].*
37	*Inappropriate data relationships among tables.*
38	*Unrealized data relationships between data members.*
39	*Not specifying NULL character properly in flat files data sources result in wrong data.*
40	*Delimiter that comes as a character in some field of the file may represent different meaning of data than the actual one.*
41	*Wrong number of delimiters in the sources (Files) causes DQ Problems.*
42	*Presence of Outliers.*
43	*Orphaned or dangling data (Data Pointing to other data which does not exists)[11]*
44	*Data and metadata mismatch.*
45	*Important entities, attributes and relationships are hidden and floating in the text fields [6][7].*
46	*Inconsistent use of special characters in various sources [6][7].*
47	*Multi-purpose fields present in data sources.*
48	*Deliberate Data entry errors (Input errors) contribute to data quality issues [4][11].*
49	*Poor data entry training causes data quality problems in data sources [11] [24].*
50	*Wrongly designed data entry forms, allowing illegibility [11] [24].*
51	*Different business rules of various data sources creates problem of data quality.*
52	*Insufficient plausibility (comparison within the data set and within time) checks in operative systems (e.g. during data input). [20]*

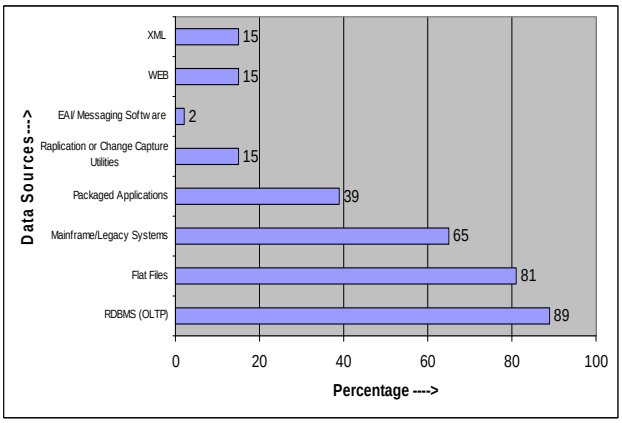
According to The Standish Group report, one of the primary causes of data migration project overruns and failures is a lack of understanding of the source data prior to data movement into some decision oriented data store such as data warehouse [17]. Especially the legacy sources are to be taken care of before we move data. So for the purpose of formulation of classification of causes of data pollution at the data source stage, we mainly identified data quality issues in following types of data sources

a) Legacy Systems
b) OLTP/ operational Systems
c) Flat/Delimited Files

And classification is confined to non multimedia (Images, Video and Audio) data only. The data sources considered is on the basis of survey conducted by Wayne Eckerson [14] in the report titled "Evaluating ETL and Data Integration Platforms". According to the survey, on average, organizations now extract data from 12 distinct data sources out which maximum is OLTP, Legacy and Flat Files by being encouraged from the report. Result of analysis is shown in Figure 5 showing the percentage of each type of data source from where companies extract data for the purpose of loading it into data warehouse.

All the causes presented in the table 1 are related to the data sources which are the feeder systems for the data warehouse. In literature problem of missing data, non standardized data and formats, and problems of data quality in legacy systems were emphasized more. One of the fundamental obstacles in the current data warehousing environment concerns the existence of inconsistent data. According to Amit Rudra and Emilie Yeo [4] as per their survey the top three reasons for data pollution in the data warehouse seems to be: Data never being fully captured, heterogeneous system integration and Lack of policy and planning from management. Our classification has presented much more number of causes of data pollution.

Figure 5:- Types of data sources Organizations extract data from. [14]



2.2.2 Causes of Data Quality Issues at Data Profiling Stage

When possible candidate data sources are identified and finalized data profiling comes in play immediately. Data profiling is the examination and

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

assessment of your source systems' data quality, integrity and consistency sometimes also called as source systems analysis. Data profiling is a fundamental, yet often ignored or given less attention as result of which data quality of the data warehouse is compromised At the beginning of a data warehouse project, as soon as a candidate data source is identified, a quick data profiling assessment should be made to provide a go/no-go decision about proceeding with the project. Table 2 is depicting the possible causes of data quality degradation at the data profiling stage of data warehousing.

Table 2: Causes of Data Quality Issues at Data Profiling Stage

Sr. No	*CAUSES OF DATA QUALITY ISSUES AT DATA PROFILING.*
1	*Insufficient data profiling of data sources is responsible for data quality issues.*
2	*Manually derived information about the data Contents in operational systems propagates poor data quality [8].*
3	*Inappropriate selection of Automated profiling tool cause data quality issues [8].*
4	*Insufficient data content analysis against external reference data causes data quality problems.*
5	*Insufficient structural analysis of the data sources in the profiling stage.*
6	*Insufficient Pattern analysis for given fields within each data store.*
7	*Insufficient column profiling, single table structural profiling, cross table structural profiling of the data sources causes data quality problems [9].*
8	*Insufficient range and distribution of values or threshold analysis for required fields.*
9	*Lack of analysis of counts like record count, sum, mode, minimum, maximum percentiles, mean and standard deviation.*
10	*Undocumented, alterations identified during profiling cause data quality problems.*
11	*Inappropriate profiling of the formats, dependencies, and values of source data*
12	*Inappropriate parsing and standardization of records and fields to a common format*
13	*Lack of identification of missing data relationships*
14	*Hand coded data profiling is likely to be incomplete and leave the data quality problems.*
15	*Unreliable and incomplete metadata of the data sources cause data quality problems [8].*
16	*User Generated SQL queries for the data profiling purpose leave the data quality problems.*
17	*Inability of evaluation of inconsistent business processes during data profiling cause data quality problems.*
18	*Inability of evaluation of data structure, data values and data relationships before data integration, propagates poor data quality.*
19	*Inability of integration between data profiling, ETL cause no proper flow of metadata which leave data quality problems.*

2.2.2 Data Quality issues at Data Staging ETL (Extraction, Transformation and Loading)

One consideration is whether data cleansing is most appropriate at the source system, during the ETL process, at the staging database, or within the data warehouse [15] [18]. A data cleaning process is executed in the data staging area in order to improve the accuracy of the data warehouse. The data staging area is the place where all 'grooming' is done on data after it is culled from the source systems. Staging and ETL phase is considered to be most crucial stage of data warehousing where maximum responsibility of data quality efforts resides. It is a prime location for validating data quality from source or auditing and tracking down data issues. There may be several reasons of data quality problems at this phase, some of the identified reasons from literature review are shown in Table 3.

Table 3: Causes of Data Quality Issues at Data Staging and ETL Phase

Sr. No	*CAUSES OF DATA QUALITY ISSUES AT DATA STAGING AND ETL PHASE.*
1	*Data warehouse architecture undertaken affects the data quality (Staging, Non Staging Architecture).*
2	*Type of staging area, relational or non relational affects the data quality.*
3	*Different business rules of various data sources creates problem of data quality.*
4	*Business rules lack currency contributes to data quality problems [4].*
5	*The inability to schedule extracts by time, interval, or event cause data quality problems.*
6	*Lack of capturing only changes in source files [24].*
7	*Lack of periodical refreshing of the integrated data storage (Data Staging area) cause data quality degradation.*
8	*Truncating the data staging area cause data quality problems because we can't get the data back to reconcile.*
9	*Disabling data integrity constraints in data staging tables cause wrong data and relationships to be extracted and hence cause data quality problems [11].*
10	*Purging of data from the Data warehouse cause data quality problems [24].*

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

11	*Hand coded ETL tools used for data warehousing lack in generating single logical meta data store, which leads to poor data quality.*
12	*Lack of centralized metadata repository leads to poor data quality.*
13	*Lack of reflection of rules established for data cleaning, into the metadata causes poor data quality.*
14	*Inappropriate logical data map prepared cause data quality issues.*
15	*Misinterpreting/Wrong implementation of the slowly changing dimensions (SCD) strategy in ETL phase causes massive data quality problems.*
16	*Inconsistent interpretation or usage of codes symbols and formats [4].*
17	*Improper extraction of data to the required fields causes data quality problems [4].*
18	*Lack of proper functioning of the extraction logic for each source system (historical and incremental loads) cause data quality problems.*
19	*Unhandled null values in ETL process cause data quality problems.*
20	*Lack of generation of data flow and data lineage documentation by the ETL process causes data quality problems.*
21	*Lack of availability of automated unit testing facility in ETL tools cause data quality problems.*
22	*Lack of error reporting, validation, and metadata updates in ETL process cause data quality problems.*
23	*Inappropriate handling of rerun strategies during ETL causes data quality problems.*
24	*Inappropriate handling of audit columns such as created date, processed date and updated date in ETL*
25	*Inappropriate ETL process of update strategy (insert/update/delete) lead to data quality problems.*
26	*Type of load strategy opted (Bulk, batch load or simple load) cause Data Quality problems. [24]*
27	*Lack of considering business rules by the transformation logic cause data quality problems.*
28	*Non standardized naming conventions of the ETL processes (Jobs, sessions, Workflows) cause data quality problems.*
29	*Wrong impact analysis of change requests on ETL cause data quality problems.*
30	*Loss of data during the ETL process (rejected records) causes data quality problems. (refused data records in the ETL process)*
31	*Poor system conversions, migration, reengineering or consolidation contribute to the data quality problems [4] [24].*
32	*The inability to restart the ETL process from checkpoints without losing data [14]*
33	*Lack of Providing internal profiling or integration to third-party data profiling and cleansing tools.[14]*
34	*Lack of automatically generating rules for ETL tools to build mappings that detect and fix data defects[14]*
35	*Inability of integrating cleansing tasks into visual workflows and diagrams[14]*
36	*Inability of enabling profiling, cleansing and ETL tools to exchange data and meta data[14]*

2.2.4 Causes of Data Quality Problems at Data Modeling (Database Schema Design) Stage.

The quality of the information depends on 3 things: (1) the quality of the data itself, (2) the quality of the application programs and (3) the quality of the database schema [19]. Design of the data warehouse greatly influences the quality of the analysis that is possible with data in it. So, special attention should be given to the issues of schema design. Some of the issues such as slowly changing dimensions, rapidly changing dimension, and multi valued dimensions etc. A flawed schema impacts negatively on information quality. Table 4 is depicting the listing of some most important causes of data quality issues at data warehouse schema designing

Table 4: Causes of Data Quality Issues at Data Warehouse Schema Modeling Phase

Sr. No	*CAUSES OF DATA QUALITY ISSUES AT DATA WAREHOUSE SCHEMA DESIGN.*
1	*Incomplete or wrong requirement analysis of the project lead to poor schema design which further casue data quality problems.*
2	*Lack of currency in business rules cause poor requirement analysis which leads to poor schema design and contribute to data quality problems.*
3	*Choice of dimensional modeling (STAR, SNOWFLAKE, FACT CONSTALLATION) schema contribute to data quality.*
4	*Late identification of slowly changing dimensions contribute to data quality problems.*
5	*Late arriving dimensions cause DQ Problems.*
6	*Multi valued dimensions cause DQ problems.*
7	*Improper selection of record granularity may lead to poor schema design and thereby affecting DQ.*
8	*Incomplete/Wrong identification of facts/dimensions, bridge tables or relationship tables or their individual relationships contribute to DQ problems.*
9	*Inability to support database schema refactoring cause data quality problems.*

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

10	*Lack of sufficient validation, and integrity rules in schema contribute to poor data quality.*

2.3 Discussion on Classification

The paper has traced out the possible causes of data quality problems at every stage of data warehousing. Talking about classification of data quality issues at data sources stage, a set of almost 52 causes is framed which contribute towards data quality of the data warehouse if not taken care of may lead to poor data quality. This stage is considered to be more vulnerable to data quality problems as the data is culled from various heterogeneous environments. The most common type of problems that are manifested in literature of data quality are: lack of standardization of data, non standardization of formats, heterogeneity of data sources, Non-Compliance of data in data sources with the standards, missing data , inconsistent data across the sources, inadequate data quality testing on individual data source and many more depicted in Table1.

Data profiling is the technical analysis of data to describe its content, consistency and structure. Profiling can be divided into following categories [22]

Pattern Analysis – Expected patterns, pattern distribution, pattern frequency and drill down analysis

Column Analysis – Cardinality, null values, ranges, minimum/maximum values, frequency distribution and various statistics.

Domain Analysis – Expected or accepted data values and ranges

Table 2 formulated the classification of causes of data quality issues pertaining to the above classification of data profiling activities. A brief set of 19 possible causes of data quality problems are identified at this stage of data profiling.

ETL and data staging is considered to be more crucial stage of data warehousing process where most of the data cleansing and scrubbing of data is done. There can be myriad of reasons at this stage which can contribute to the data quality problems. A common mistake is to write custom programs to perform the extraction, transformation, and load functions. Writing an ETL program by hand may seem to be a viable option because the program does not appear to be too complex and programmers are available [23]. However, there are serious problems with hand-coded ETL programs. It give rise to problems like metadata is not generated automatically by hand-generated ETL programs, To keep up with the high volume of changes initiated by end users, hand-written ETL programs have to be constantly modified and in many cases rewritten, and much more problems can be there which may contribute to the data quality problems. Some of the problems are related to the capabilities of off self ETL tools use for the purpose of data warehousing and some are the traditional problems related to this phase. A brief set of 36 such issues have been identified at this stage and are mentioned in table 3.

Data Warehouse schema design and modeling is also considered to be vulnerable for contributing towards data quality problems. A flawed schema impacts negatively on information quality. Taking care of some of the factors related to data quality while designing the schema for warehouse would really help in achieving quality data in it. Table 4 depicts a set of 10 causes which if not taken care of while you design schema for warehouse would really deteriorate the data quality.

2.4 Conclusions

Today, to the best of our knowledge, no comprehensive & descriptive classification of causes of data quality problems exists. In this paper we have attempted to collect all possible causes of data quality problems that may exist at all the phases of data warehouse. Our objective was to put forth such a descriptive classification which covers all the phases of data warehousing which can impact the 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, even some researchers have attempted to provide brief set of issues of data quality problems as well. But none of the research has attempted to think on near possible set of causes of data quality problems at all the phases at one attempt. Our classification of causes will really help the data warehouse practitioners, implementers and researchers for taking care of these issues before moving ahead with each phase of data warehousing. It would also be helpful for the vendors and those who are involved in development of data quality tools so as to incorporate changes in their tools to overcome the problems highlighted in classifications

2.5 Future work

Each item of the classification shown in table 1,2,3 and 4, will be converted into a item of the research instrument such as questionnaire and will be empirically tested by collecting views about these items from the data warehouse practitioners, appropriately.

Acknowledgements

Much of the information in this paper was derived from extended conversations with Subject Matter Experts

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

(SME), data warehouse practitioners, Data Quality Stewards of various IT companies of India and a through literature review of data quality issues. The authors gratefully acknowledge the time investments in this project that were generously provided by Mr.Saurabh Chopra, SME, AMDOCS, Pune India, Ms. Ravneet Bhatia and Ms. Rosy Choudhary Software Engineer (Data Warehousing Wing). ebusinessware India Pvt. Ltd. , Mr. Gursimran Singh Sr.Software Engineer (DBA) IBM India Pvt Ltd, Gurgaon, India. Special Thanks to people from abroad Mr. Girish Butaney and Mr. Nirlap Vora, Mr. Kerri Patel of SAS, Mr. Won Kim, Mr. Thilini Ariyachandra Assistant Professor of Management Information Systems Williams College of Business, Xavier University, Professor Yair Wand Head at the University of British Columbia, Faculty of Commerce and Business Administration, in Vancouver. *Jean-Pierre Dijcks* Principal Product Manager. Oracle Warehouse Builder. Oracle Corporation.

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

[4] Amit Rudra and Emilie Yeo (1999) "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd Hawaii International Conference on System Sciences – 1999

[5] Jesús Bisbal et all (1999) "Legacy Information Systems: Issues and Directions", *IEEE Software September/ October 1999*

[6] Scott W. Ambler (2001) "Challenges with legacy data: Knowing your data enemy is the first step in overcoming it", Practice Leader, Agile Development, Rational Methods Group, IBM, 01 Jul 2001.

[7] Scott W. Ambler "The Joy of Legacy Data" available at: http://www.agiledata.org/essays/legacyDatabases.html

[8] Wayne Eckerson, *"Data Profiling: A Tool Worth Buying (Really!)" ,Information Management Magazine, June 1, 2004 available http://wwwInformation-management.com/issues/20040601/1003990-1.html*

[9] "Boosting Data Quality for Business Success", INFORMATICA White Paper, 2006

[10] Federico Zoufaly *"Issues and Challenges Facing Legacy Systems ", available at http://www.developer.com/mgmt/article.php/11085_1492531*

[11] Won Kim et al (2002)- *"A Taxonomy of Dirty Data " Kluwer Academic Publishers 2002*

[12] Erhard Rahm & Hong Hai Do (2003) *"Data Cleaning: Problems and Current Approaches "*, available at: http://homepages.inf.ed.ac.uk/wenfei/tdd/reading/cleaning.pdf.

[13] HarteHanks *"Solving the source data problems with automated data profiling" Trillium Software, available at www.trilliumsoftware.com*

[14] Wayne Eckerson & Colin White (2003) *" Evaluating ETL and Data Integration Platforms" TDWI report series*

[15] 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 .

[16] Mike(2009), " "The problem of dirty data" available at http://www.articlesbase.com/databases-articles/the-problem-of-dirty-data-1111299.html

[17] The Standish Group (1999), "Migrate Headaches," available at *www.it-cortex.com/start_failure_rate.htm*

[18] Ralaph Kimball, The Data Warehouse ETL Toolkit, Wiley India (P) Ltd (2004),

[19] Tech Notes (2008), Why Data Warehouse Projects Fail: Using Schema Examination Tools to Ensure Information Quality, Schema Compliance, and Project Success. Embarcadero Technologies. Available at www.embarcadero.com.

[20] Markus Helfert, Gregor Zellner, Carlos Sousa, "Data Quality Problems and Proactive Data Quality Management in Data-Warehouse-Systems",

[21] David Loshin , "The Data Quality Business Case: Projecting Return on Investment", Informatica White paper. Available at :

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

http://www.melissadata.com/enews/articles/1007/2.htm

[22] Amol Shrivastav,Mohit Bhaduria, Harsha Rajwanshi (2008), " Data Warehouse and Quality Issues", available at http://www.scribd.com/doc/9986531/Data-Warehouse-and-Quality-Issues

[23] Ahimanikya Satapathy, "Building an ETL Tool", Sun Microsystems, Available at : http://wiki.open-esb.java.net/attach/ETLSE/ETLIntroduction.pdf

[24] Arkedy Maydanxhik (2007), " Causes of Data Quality Problems", Data Quality Assessment, Techniques Publications LLC. Available at http://media.techtarget.com/searchDataManagement/downloads/Data_Quality_Assessment_-_Chapter_1.pdf

Ranjit Singh is a research fellow with university college of Engineering & Technology (UCoE), Punjabi University Patiala, Punjab (INDIA). He is currently working on his PhD, "Development of Data Quality Assurance Model (DQAM) for Data Warehouse and its Impact on Business Intelligence". He has been awarded his Master of Computer Applications (M.C.A.) degree in 2002 from G.N.D.U, Amritsar, Punjab, India, and Master of Philosophy (M.Phil.) in Computer Science in 2006 from M.K.U. Madurai, Tamil Nadu, India. Presently author is working as senior lecturer in department of computer applications at Apeejay Institute of Management, Jalandhar, Punjab, India. He has authored 4 books in the area of computer applications and more than 6 research papers in national level journals and presented more than 10 papers in conference seminars at national level.

Dr. Kawaljeet Singh is working as Director, University Computer Center (UCC) at Punjabi University Patiala, Punjab, Inida. He has earned his Master of Computer Applications (MCA) in year 1988 and Doctorate of Philosophy (Ph.D.) in year 2001 in computer science from Thaper Institute of Engineering & Technology (T.I.E.T.), Deemed University, Patiala Punjab, India. He is having near about two decades of professional, administrative and academic career. He has served for more than a decade at various positions in Guru Nanak Dev University (GNDU) Amritsar, Punjab, India such as Reader & Head in Department of Computer Science and Engineering (DCSCE) , Amritsar campus, Professor and Head of the DCSCE & Electronics in Regional Campus of GNDU Jalandhar. He has also headed many responsible positions such as Dean & Chairman of Research Degree of the Committee of Faculty of Engineering & Technology of G.N.D.U. He has also co-authored 3 text books and published more than 32 research papers, abstracts, key notes and articles in national and international journals, conferences, seminars and workshops.