IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. IV (Jan. 2014), PP 18-24
www.iosrjournals.org

Data Quality in Data warehouse: problems and solution

*Rahul Kumar Pandey,
PhD Scholar, Surguja university (Chhattisgarh),India

Abstract: *In recent years, corporate scandals, regulatory changes, and the collapse of major financial institutions have brought much warranted attention to the quality of enterprise data if we can better understand the problems of quality issues, then we can develop a plan of action to address the problem that is both proactive and strategic. Each instance of a quality issue presents challenges in both identifying where problems exist and in quantifying the extent of the problems. Quantifying the issues is important in order to determine where our efforts should be focused. It is reported that more than $2 billion of U.S. federal loan money had been lost because of poor data quality at a single agency. It also reported that manufacturing companies spent over 25% of their sales on wasteful practices. 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 along with their possible solution. problems in different phase of data warehouse i.e.; data sources, data integration & data profiling, Data staging and ETL, data warehouse modeling & schema design are discussed in this paper.*

The 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 give some classification of these causes as well as solution for improving data quality through Statistical Process Control (SPC),Quality engineering management . etc I 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 company on India. I 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), Statistical Process Control (SPC),ETL, Data Staging, Data Warehouse*

I. Data Warehouse:

Data Warehouse (DW) is a collection of technologies aimed at enabling the knowledge worker (executive,manager,analyst, etc) to make better and faster decisions. Many researchers and practitioners share that a data warehouse architecture can be formally understood as layers of materialized views on top of each other. A data warehouse architecture exhibits various layers of data in which data from one layer are derived from data of the lower layer.

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.

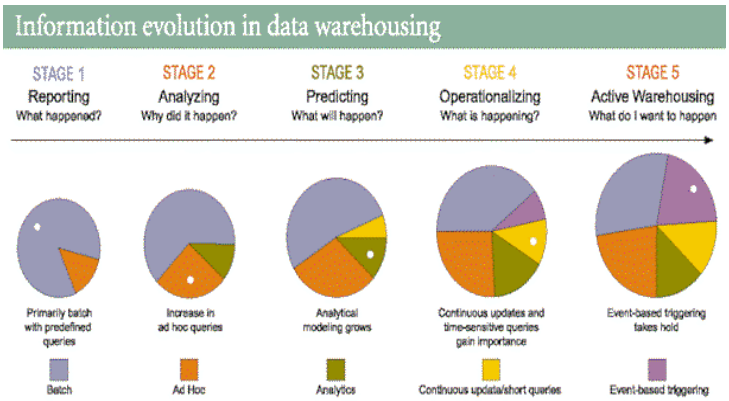

fig: five stages of data warehouse

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 processable 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 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 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.

Data Quality Issues at Data Sources:

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,

Table1: causes of data quality at Data source

Sr.No	***CAUSES OF DATA QUALITY PROBLEMS AT DATA SOURCES***
1	*Wrong information entered into source system*
2	*As time and proximity from the source increase, the chances for getting correct data decrease*
3	*In adequate knowledge of interdependencie among data sources incorporate DQ problems.*
4	*Inability to cope with ageing data contribute to data quality problems*
5	*Varying timeliness of data sources*
6	*Complex Data warehouse*
7	*Unexpected changes in source systems cause DQProblems*
8	*System fields designed to allow free forms (Field not having adequate length).*

9	*Missing values in data sources*
10	*Additional columns*
11	*Use of different representation formats in data sources*
12	*Non-Compliance of data in data sources with the Standards*
13	*Failure to update all replicas of data causes DQ Problems.*
14	*Approximations or surrogates used in data*
15	*Different encoding formats (ASCII, EBCDIC,….)*
16	*Lack of business ownership, policy and planning of the entire enterprise data contribute to data quality problems.*

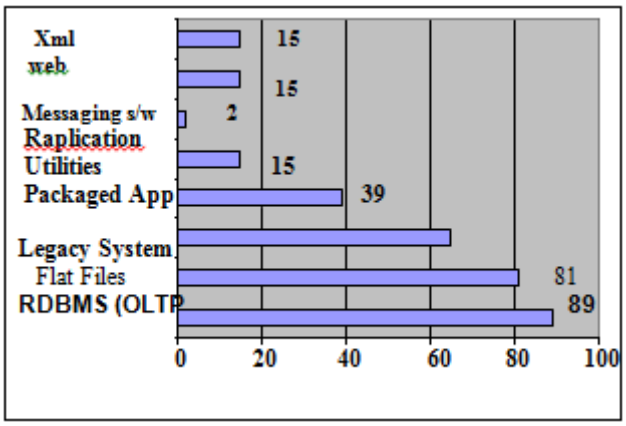

Fig:1 Types of data sources Organizations extract data from.

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

Table 2:causes of data quality at Data profiling

Sr.No	***CAUSES OF DATA QUALITY PROBLEMS AT DATA PROFILING***
1	*Insufficient range and distribution of values or threshold analysis for required fields.*
2	*Unreliable and incomplete metadata of data source*
3	*UserGeneratedSQLqueriesforthedataprofilingpurpose leave the data quality problems.*
4	*Inability of evaluation of inconsistent business processes during data profiling cause data quality problems.*
5	*Inability of evaluation of data structure, data values and data relationships before data integration ,propagates poor data quality*
6	*Inability of integration between Data profiling and ETL causes Data quality problem*
7	*Inappropriate selection of Automated profiling tool cause data quality issues*
8	*Insufficient data content analysis against external reference data causes data quality problems.*
9	*Insufficient structural analysis of the data sources in the profiling stage.*
10	*Insufficient Pattern analysis for given fields within each data store*

Data Quality issue at D ata Staging ETL(Extra action), Transformation and Loading)

One consideration is whether data cleansing is most appropriate a the source system, during the ETL process, at the staging database,or within the data warehouse..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. Some of the identified area from literature review are shown in Table

Table3: 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*
5	*Business rules lack currency contributes to data quality problems*
6	*Lack of capturing only changes in source files*
7	*Lack of periodical re freshing of the integrated data storage (Data Staging area) cause data quality degradation*
8	*Truncating the data staging area cause data quality Problems*
9	*Disabling data integrity constraints in data staging tables cause wrong data and relationships to be extracted and hence cause data quality problem*
10	*Purging of data from the Data warehouse \cause data quality problem*
11	*The inability to restart the ETL process from checkpoints without losing data*
12	*Lack of Providing internal profiling or integration to third-party data profiling and cleansing tools.[*
13	*Lack of automatically generating rules for ETL tools to build mapping s that detect and fix data defects*
14	*Inability of integrating cleansing tasks into visual work flows and diagrams*
15	*Unhandled null values causes data quality problem*
16	*Lack of automated unit testing facility causes data quality problem*

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 init. 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 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	*CAUSESOF DATA QUALITY ISSUES AT DATA WAREHOUSE SCHEM A DESIGN.*
1	*Incomplete or wrong requirement analysis of the project lea d to poor schema design which further cause 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,FACTCONSTALLATION) schema contribute to data quality.*
4	*Late identification of slowly changing dimensions contribute to data qualit y 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. Problem*
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*

Suggestion for improving Data quality:

Enterprise Architecture

Creating and maintaining enterprise architecture (EA) is a effective method for controlling data redundancies as well as process redundancies, and thereby reducing the anomalies and inconsistencies that are inherently produced by uncontrolled redundancies. EA is comprised of models that describe an organization in terms of its business architecture (business functions, business processes, business data, and so on) and technical architecture (applications, databases, and so on). Purpose of these models is to describe the actual business in which the organization engages. EA is applicable to all organizations, large and small. Because EA models are best built incrementally, one project at a time, it is appropriate to develop EA models on DW and BI projects, as well as on projects that simply solve departmental challenges.

Data Quality Improvement Process

In addition to applying enterprise-wide data quality disciplines, creating an enterprise data model, and documenting metadata, the data quality group should develop their own data quality improvement process. At the highest level, this process must address the six major components shown in figure These components are:

.Assess—Every improvement cycle starts with an assessment. This can either be an initial enterprise-wide data quality assessment, a system-by-system data quality assessment, or a department-by-department data quality assessment. When performing the assessment, do not limit your efforts to profiling the data and collecting statistics on data defects. Analyze the entire data entry or data manipulation process to find the root causes of errors and to find process improvement opportunities.

Another type of assessment is a periodic data audit. This type of assessment is usually limited to one file or one database at a time. It involves data profiling as well as manual validation of data values against the documented data domains (valid data values)

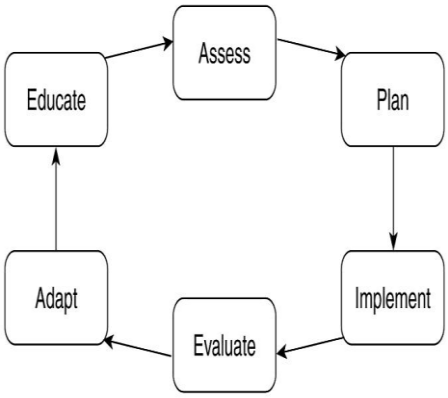

Fig 2: Data quality improvement cycle

- **Plan**—After opportunities for improvement have been defined, the improvements should be analyzed, prioritized, approved, funded, staffed, and scheduled. Not all improvements have the same payback and not all improvements are practical or even feasible. An impact analysis should determine which improvements have the most far-reaching benefits. After improvement projects have been prioritized, approved, and funded, they should be staffed and scheduled.

.Implement—In some cases, the data quality group can implement the approved improvements, but in many cases, other staff members from both the business side and IT will be required. For example, a decision might have been made that an overloaded column (a column containing data values describing multiple attributes) should be separated in a database. That would involve the business people who are currently accessing the database, the database administrators who are maintaining it, and the developers whose programs are accessing it.

Evaluate—the best ideas sometimes backfire. Although some impact analysis will have been performed during planning, occasionally an adverse impact will be overlooked. Or worse, the implemented improvement might have inadvertently created a new problem. It is therefore advisable to monitor the implemented improvements and evaluate their effectiveness. If deemed necessary, an improvement can be reversed.

.Adapt—hopefully, most improvements do not have to be reversed, but some may have to be modified before announcing them to the entire organization or before turning them into new standards, guidelines, or procedures.
.Educate—The final step is to disseminate information about the new improvement process just implemented. Depending on the scope of the change, education can be accomplished through classroom training, computer-based training, an announcement on the organization's intranet website, an internal newsletter, or simple e-mail notification.

Data Quality Disciplines:

Organizations have a number of data quality disciplines at their disposal, but rarely will they implement all disciplines at once because improving data quality is a process and not an event .
Figure shows the common data quality improvement activities in each of the five data quality maturity levels based on Larry English's adaptation of the capability model (CMM) to data quality. The five levels are:

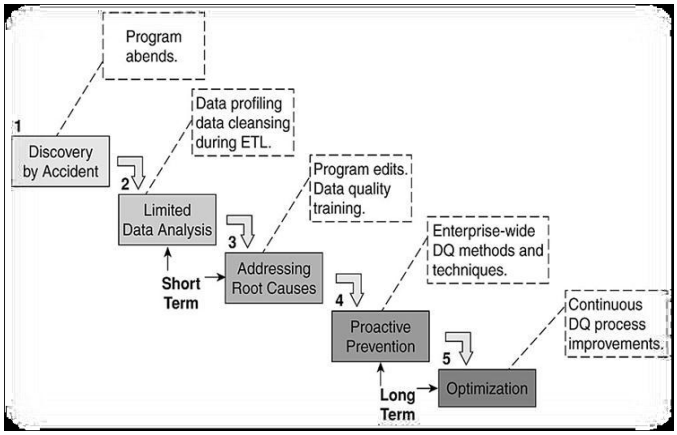

Fig 3 data quality improvement activities

Level 1: Uncertainty—At Level 1, the organization is stumbling over data defects as its programs abend (crash) or its information consumers complain. There is no proactive data quality improvement process, no data quality group, and no funding. The organization denies any serious data quality problems and considers data analysis a waste of time. Or the CIO is ready to retire and doesn't want anything to disrupt it. Basically, the organization is asleep and doesn't want to be awakened.

Level 2: Awakening—At Level 2, the organization performs some limited data analysis and data correction activities, such as data profiling and data cleansing. There still is no enterprise-wide support for data quality improvement, no data quality group, and no funding. However, a few isolated individuals acknowledge their dirty data and want to incorporate data quality disciplines in their projects. These individuals can be data administrators, database administrators, developers, or business people.

Level 3: Enlightenment—At Level 3, the organization starts to address the root causes of its dirty data through program edits and data quality training. A data quality group is created and funding for data quality improvement projects is available. The data quality group immediately performs an enterprise-wide data quality assessment of their critical files and databases, and prioritizes the data quality improvement activities. This group also institutes several data quality disciplines and launches a comprehensive data quality training program across the organization.

Level 4: Wisdom—At Level 4, the organization proactively works on preventing future data defects by adding more data quality disciplines to its data quality improvement program. Managers across the organization accept personal responsibility for data quality. The data quality group has been moved under a chief officer—either the CIO, COO, CFO, or a new position, such as a chief knowledge officer (CKO). Metrics are in place to measure the number of data defects produced by staff, and these metrics are considered in the staff's job performance appraisals. Incentives for improving data quality have replaced incentives for cranking out systems at the speed of light.
Level 5: Certainty—At Level 5, the organization is in an optimization cycle by continuously monitoring and improving its data defect prevention processes. Data quality is an integral part of all business processes. Every job description requires attention to data quality, reporting of data defects, determining the root causes,

improving the affected data quality processes to eliminate the root causes, and monitoring the effects of the improvement. Basically, the culture of the organization has changed.

II. 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

This paper is helpful for 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

Future work

Each item of the classification 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.

References

[1] A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing ,R Singh ,Dr k Singh.
[2] http://www.informit.com/articles.
[3] "TDWI Data Cleansing: Delivering High-Quality Warehouse Data."
[4] Thibodeau, Patrick. "Data Problems Thwart Effort to Count
[5] AmolShrivastav,MohitBhaduria, Harsha Rajwanshi (2008), " Data Warehouse and Quality Issues", Warehouse-and-Quality-Issues.
[6] Ahimanikya Satapathy, "Building an ETL Tool", Sun Microsystems