An Analysis and Study on Data Quality in Data Warehousing

1st Parthik R. Patel
[1]Assistant Professor
[1]Shree Swaminarayan College of Computer Science, Bhavnagar, India
2nd Parag B. Makwana
[2]Assistant Professor
[2]Shree Swaminarayan College of Computer Science, Bhavnagar, India
3rd Madhav K. Dave
[3]Assistant Professor
[3]Shree Swaminarayan College of Computer Science, Bhavnagar, India

Abstract - Data quality in a data warehouse is not just the quality of individual data items but the quality of the full, integrated system as a whole. data quality dimensions and use the list to recognize and measure the data quality in the systems that feed your data warehouse. The data item is exactly fit for the purpose for which the business users have defined it.

Index Terms – Data Accuracy, Domain Integrity, Consistency, Redundancy, Data Anomaly.

I. Introduction of Data Quality

Data quality define boosts confidence in decision making, enables better customer service, increases opportunity to add better value to the services ,reduces risk from disastrous decisions, reduces costs, especially of marketing campaigns, enhances strategic decision making, improves productivity by streamlining processes, and avoids compounding effects of data contamination.

data quality dimensions and use the list to recognize and measure the data quality in the systems that feed your data warehouse.

Accuracy. The value stored in the system for a data element is the right value for that occurrence of the data element. If you have a customer name and an address stored in a record, then the address is the correct address for the customer with that name. If you find the quantity ordered as 1000 units in the record for order number 12345678, then that quantity is the accurate quantity for that order.

Domain Integrity. The data value of an attribute falls in the range of allowable, de- fined values. The common example is the allowable values being "male" and "female" for the gender data element.

Data Type. Value for a data attribute is actually stored as the data type defined for that attribute. When the data type of the store name field is defined as "text," all in- stances of that field contain the store name shown in textual format and not numeric codes.

Consistency. The form and content of a data field is the same across multiple source systems. If the product code for product ABC in one system is 1234, then the code for this product must be 1234 in every source system.

Redundancy. The same data must not be stored in more than one place in a system. If, for reasons of efficiency, a data element is intentionally stored in more than one place in a system, then the redundancy must be clearly identified.

Completeness. There are no missing values for a given attribute in the system. For example, in a customer file, there is a valid value for the "state" field. In the file for order details, every detail record for an order is completely filled.

Duplication. Duplication of records in a system is completely resolved. If the product file is known to have duplicate records, then all the duplicate records for each product are identified and a cross-reference created.

Conformance to Business Rules. The values of each data item adhere to prescribed business rules. In an auction system, the hammer or sale price cannot be less than the reserve price. In a bank loan system, the loan balance must always be positive or
zero.

Structural Definiteness. Wherever a data item can naturally be structured into individual components, the item must contain this well-defined structure. For example, an individual's name naturally divides into first name, middle initial, and last name. Values for names of individuals must be stored as first name, middle initial, and lastname. This characteristic of data quality simplifies enforcement of standards and re-duces missing values.

II. BENEFITS OF IMPROVED DATA QUALITY

Analysis with Timely Information: Suppose a large retail chain is running daily promotions of many types in most of its 200 stores in the country. This is a major seasonal campaign. Promotion is one of the dimensions stored in the data warehouse. The marketing department wants to run various analyses using promotion as the primary dimension to monitor and tune the promotions as the season progresses. It is critical for the department to perform the analyses every day.

Better Customer Service: The benefit of accurate and complete information for customer service cannot be overemphasized. Let us say the customer service representative at a large bank receives a call. The customer at the other end of the line wants to talk about the service charge on his checking account. The bank customer service representative notices a balance of $27.38 in the customer's checking account. Why is he making a big fuss about the service charge with almost nothing in the account? But let us say the customer service representative clicks on the customer's other accounts and finds that the WHY IS DATA QUALITY CRITICAL? 295 customer has $35,000 in his savings accounts and CDs worth more than $120,000. How do you think the customer service representative will answer the call? With respect, of course. Complete and accurate information improves customer service tremendously.

Newer Opportunities : Quality data in a data warehouse is a great boon for marketing. It opens the doors to immense opportunities to cross-sell across product lines and departments. The users can select the buyers of one product and determine all the other products that are likely to be purchased by them. Marketing departments can conduct well-targeted campaigns. This is just one example of the numerous opportunities that are made possible by quality data. On the other hand, if the data is of inferior quality, the campaigns will be failures.

Reduced Costs and Risks: What are some of the risks of poor data quality? The obvious risk is strategic decisions that could lead to disastrous consequences. Other risks include wasted time, malfunction of processes and systems, and sometimes even legal action by customers and business partners. One area where quality data reduces costs is in mailings to customers, especially in marketing campaigns. If the addresses are incomplete, inaccurate, or duplicate, most of the mailings are wasted.

Improved Productivity :Users get an enterprise-wide view of information from the data warehouse. This is a primary goal of the data warehouse. In areas where a corporate-wide view of information naturally enables the streamlining of processes and operations, you will see productivity gains. For example, a company-wide view of purchasing pat-terns in a large department store can result in better purchasing procedures and strategies

III. DATA QUALITY CHALLENGES

Sources of Data Pollution

In order to come up with a good strategy for cleansing the data, it will be worthwhile to review a list of common sources of data pollution. Why does data get corrupted in the source systems? Study the following list of data pollution sources against the background of what data quality really is.

System conversions

Trace the evolution of order processing in any company. The company must have started with a file-oriented order entry system in the early 1970s; orders were entered into flat files or indexed files. There was not much stock verification or customer credit verification during the entry of the order. Reports and hard-copy printouts were used to continue with the process of executing the orders. Then this system must have been converted into an online order entry system with VSAM files and IBM's CICS as the online processing monitor. The next conversion must have been to a hierarchical database system. Perhaps that is where your order processing system still remains—as a legacy application.

Data aging

We have already dealt with data aging when we reviewed how over the course of many years the values in the product code fields could have decayed. The older values lose their meaning and significance. If many of your source systems are old legacy systems, pay special attention to the possibility of aged data in those systems.

Heterogeneous system integration

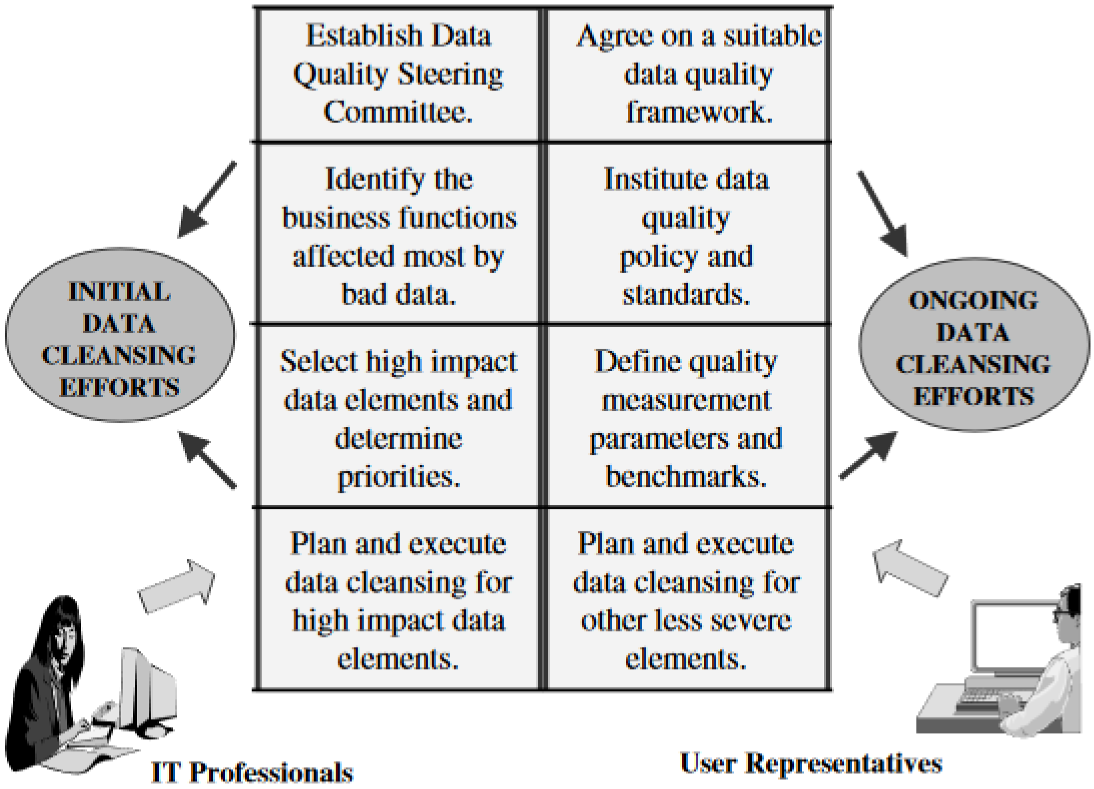
The more heterogeneous and disparate your source systems are, the stronger is the possibility of corrupted data. In such a scenario, data inconsistency is a common problem. Consider the sources for each of your dimension tables and the fact table. If the sources for one table are several heterogeneous systems, be cautious about the quality of data coming into the data warehouse from these systems.

Poor database design

Good database design based on sound principles reduces the introduction of errors. DBMSs provide for field editing. RDBMSs enable verification of the conformance to business rules through triggers and stored procedures. Adhering to entity integrity and referential integrity rules prevents some kinds of data pollution.

IV. Data Quality Framework & participants and roles

You have to contend with so many types of data pollution. You need to make various decisions to embark on the cleansing of data. You must dig into the sources of possible data corruption and determine the pollution. Most companies serious about data quality pull all these factors together and establish a data quality framework. Essentially, the framework provides a basis for launching data quality initiatives. It embodies a systematic plan for action. The framework identifies the players, their roles, and responsibilities. In short, the framework guides the data quality improvement effort


Establish Data Quality Steering Committee.
Agree on a suitable data quality framework.
Identify the business functions affected most by bad data.
Institute data quality policy and standards.
INITIAL DATA CLEANSING EFFORTS
ONGOING DATA CLEANSING EFFORTS
Select high impact data elements and determine priorities.
Define quality measurement parameters and benchmarks.
Plan and execute data cleansing for high impact data elements.
Plan and execute data cleansing for other less severe elements.
IT Professionals
User Representatives

[Data Quality Framework]

Figure shows the participants in the data quality initiatives. These persons represent the user departments and IT. The participants serve on the data quality team in specific roles. Listed below are the suggested responsibilities for the roles: Data Consumer. Uses the data warehouse for queries, reports, and analysis. Establishes the acceptable levels of data quality.

Data Producer: Responsible for the quality of data input into the source systems.

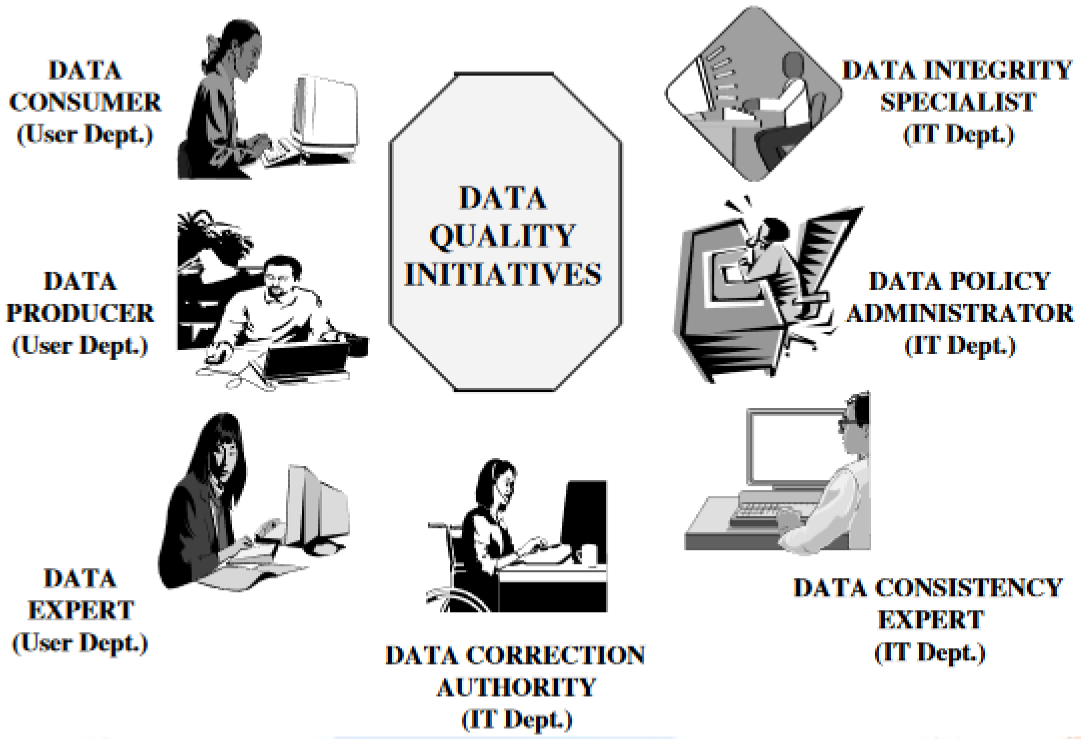
Data Expert: Expert in the subject matter and the data itself of the source systems. Responsible for identifying pollution in the source systems.

Data Policy Administrator: Ultimately responsible for resolving data corruption as data is transformed and moved into the data warehouse.

Data Integrity Specialist. Responsible for ensuring that the data in the source systems conforms to the business rules.

Data Correction Authority. Responsible for actually applying the data cleansing techniques through the use of tools or in-house programs.

Data Consistency Expert. Responsible for ensuring that all data within the data warehouse (various data marts) are fully synchronized.


DATA CONSUMER (User Dept.)
DATA PRODUCER (User Dept.)
DATA EXPERT (User Dept.)
DATA QUALITY INITIATIVES
DATA CORRECTION AUTHORITY (IT Dept.)
DATA INTEGRITY SPECIALIST (IT Dept.)
DATA POLICY ADMINISTRATOR (IT Dept.)
DATA CONSISTENCY EXPERT (IT Dept.)

[Data Participants and Roles]

V. REFERENCES

[1] Paulraj Ponniah.. Data Warehousing Fundamentals
[2] Ralph kimbal . The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling