



## **Data Quality and its importance in A corporate Data Warehouse**

### **Abstraction**

The DOT Group has taken the best of breed approach to Data Warehousing technology. This approach consists of six key components: extract and manage, organise and maintain, and deliver and exploit. Data warehouse management deals with the extraction, loading, transformation, and integration of data.

Data Quality is a key element of data warehousing and to the overall value of decisions made using the database as an information tool. This paper will discuss the common data Issues, the history of data cleansing, the approaches, the benefits to the business and the future methodologies of data cleansing.

### **INDEX**

#### **The Importance of Data Quality**

##### **Common Data Issues**

Missing or Inaccurate Data

Human Error

Flags

Homonyms

Lack of Data Standards

##### **History of Data Cleansing**

The Manual Approach

First Generation Tools

##### **The Current Approach to Data Analysis and Cleansing**

Platform Independent and Portable

Adaptable and Customisable

Reusability

Ease of use

##### **Benefits to the Business**

##### **Summary**

### **The Importance of Data Quality**

**TOP**

It is a known fact that in today's commercial environment key business decisions are based upon the information that is held in company's databases. Reporting tools such as Business Objects and Cognos let us access this information in a condensed or formatted

view, letting managers understand their business quickly and efficiently. This methodology should lead to improved decision-making due to the accurate information available to management in regard to the company or department's status. This however in our experience is not always the case. The information extracted from Databases is primarily based upon the data that is entered at its lowest level of granularity. If this data contains anomalies or errors, then the decisions that are reached from it will be flawed due to inaccurate result sets; as the old adage goes 'rubbish in, rubbish out'.

It is also acknowledged that the size and complexity of databases is increasing exponentially, two things result in this occurring:

- We rely more and more on our databases, at times becoming fully dependent on them.
- The number of errors in data increases and the errors become harder to trace.

(Quality Unbound; K.Parsaye and M.Chignell, 1998)

This led to the concept of Information architecture which spawned projects such as Operational Data Stores (ODS), Data Warehousing and Data Marts (Achieving Enterprise Data Quality; Trillium Software). Companies believed by implementing such projects data quality would improve as part of the Extraction, Translation and loading (ETL) phases into the new data stores, this however does not remove the errors in the data.

Data Quality is a key element to the overall value of decisions made using the database as an information tool. This paper will discuss the common data Issues, the history of data cleansing, the approaches, and the benefits to the business.

## **Common Data Issues:**

[TOP](#)

Almost everyone who is involved in databases is aware that there are errors present in their data. These can arise from a number of different factors including invalid data entry e.g. human error, default values not being appropriate for a given field or date entries being used as a flag that could lead to reports being misleading to the interpreter. The origin of such errors are easy to understand and will be discussed in the following chapter, it is the removal and prevention of such errors that are hard to eliminate.

### **Missing or Inaccurate Data:**

In our experience missing or inaccurate data accounts for a large proportion of data errors in large corporate warehouses. Missing data can be detected easily, and by applying conditions to important fields, prevented.

However, if columns contain data that is valid for a particular field this may not still meet the requirements/accuracy for the overlying business. For example, the address "96 Whittington Road, Westleigh" may be syntactically correct, but there may be many

duplicates of this address around the country. Likewise, the name "J Smith" is correct but without a Christian name or title, finding the relevant linked data to that person would be impossible. Conditions therefore would have to be applied to these fields such as a town or postcode or title and Christian name to make them complete.

### **Human Error:**

Again this is another common factor for error occurrence in databases. Human error can generally account for misspellings, typographical errors, out of range values or incorrect data types (Teaming up to manage data; N.Zurell). Conditions on columns and validation routines can easily pick up out of range data or incorrect data, but detecting misspellings or topographical errors is more complicated and requires quality checks to be in place.

### **Flags:**

Phony data is often used as a flag to indicate various states for a particular record, for example that that record is no longer valid. Dates outside date ranges are frequently used for indicators, such as '01/01/2010', this would not represent for example an order date but an indication of the status of a record. This is obviously a valid date and can be entered into a date field, and used by one department as an indicator flag. But for other departments or higher level reporting this inappropriate date may mislead or create inaccurate figures.

### **Homonyms:**

The English language contains many words and abbreviations with identical spellings that have multiple, and often unrelated or conflicting meanings, and relies on the context of usage to determine the correct meaning. Improper interpretation of the context in which the homonym was used can have a significant impact on data accuracy (Achieving Enterprise Data Quality; Trillium Software).

A good example of this is in the many ways we may interpret the abbreviation 'St', as shown below:

Elizabeth B. **St.** James, MD.  
In trust for  
Mary Church  
**St.** Catherine's Church  
11 High **St.**  
**St.** Petersburg, FL 33708

### **Lack of Data Standards:**

Data entry into a system is usually accounted for by multiple personnel and sometimes by multiple departments. This means that responsibilities are spread between many people all with different interpretations of the format of the required field. For example an IT equipment can be entered in different ways such as; a 'PC', 'Computer', or

'Personal Computer'. All of these are valid, but would these be picked up by queries or reporting tools?.

## **History of Data Cleansing**

[TOP](#)

Since Databases have grown tremendously in size, and data Warehousing Architecture has been implemented by many organisations, practitioners have known that some form of checking and cleansing of data has been needed.

### **Manual Approach:**

The initial response was to design custom built applications consisting of manually written code to reengineer data errors. These programmes were designed to sit between the source and target systems and run in conjunction with the Extraction and Loading phase during Data Warehouse Implementation.

This is a very manual process and requires hundreds of lines of code to be written, updated and maintained throughout the life span of data manipulation. Additional programmes would also be needed for each source system added to the warehousing project, of course at an added cost to the business.

### **First Generation Tools:**

This led to the first generation of data cleansing tools. These analysed anomalies, defects and invalid data before editing and fixing the data whilst parsing into the target system. These programmes relied on the paradigm of creating streams of programmatic logic to check and fix the data albeit through a GUI front end.

However, this again led to complex unmanageable applications that required large amounts of time and resources that couldn't be implemented across an enterprise wide solution. You can liken this to the equivalent in the manufacturing industry - the process of scrap and rework; fixing problems at the end of the production line (Achieving Enterprise Data Quality; Trillium Software). This is obviously a wasteful process whether in the manufacturing industry or in Data warehousing and has led to Second-generation tools utilising an Enterprise wide approach.

## **The Current Approach to Data Analysis and Cleansing**

[TOP](#)

The next generation approach combines business rules, validation techniques and procedures to create an enterprise wide solution to cleanse both at the source and during data movement. This provides the most efficient and cost effective way to reduce and eliminate data errors.

There are numerous tools which now fit this format, each of which follow the same rules

for adopting this methodology as discussed above.

- They are platform independent and portable. Corporate environments today will contain many different platforms, such as old mainframe and legacy systems to UNIX and NT Servers. Maintained on these platforms will be numerous different database applications holding data in various formats. A data-cleansing tool must understand and be able to communicate with all of these efficiently.
- Tools must also be adaptable and customisable. Any data-cleansing programme requires the ability to handle various rule sets, often dependent on the industry the data is from, and conditions set upon individual fields that may often change.
- As discussed previously, reusability is one of the major advantages with second-generation tools. If processes have been designed and implemented, they can be adapted and used again for another aspect of the data cleansing project. This makes the product more cost effective, by not only introducing reusability but also allowing increased employee efficiency in the area of maintenance.
- Finally ease of use and a powerful engine are the last criteria. An intuitive interface combined with a GUI front end makes for rapid development, efficient handovers and process understanding for all types of user. Behind the front end a powerful software solution is need to process large volumes of data quickly.

## **Benefits to the Business**

[TOP](#)

Projects to cleanse data are often implemented when a 'crisis' occurs, or a key project is in danger of failure. It is often the case even then that only known data issues are investigated and tackled rather than activating a department or company wide approach to data standards. By implementing a Data standards project in an enterprise wide approach using rule based procedures means a variety of improvements can be gained to your overall business, these are discussed below in more detail.

Instigating business rules to the adopted approach to data cleansing will gain companies additional benefits. By making business rules viable and overt businesses accomplish 2 things.

- Explicit understanding of what data exists and the state that it is in.
- Simplified maintenance in both the short and long term.

(Achieving Enterprise Data Quality; Trillium Software)

### **Benefits:**

1. The single most important reason for improving data quality within an organisation is to improve customer relationships. By keeping accurate data on customers and the business they are supplying you with an improved customer service can be initiated and a better understanding of their needs is to be gained. As it has been said, 'it is much cheaper to keep existing customers, than to attract new ones'.

2. The data that is interrogated by reporting tools on a daily basis can be completely relied upon. It is even possible to give end-users detailed accuracy figures on the reliability of the data, reducing assumptions and guess work involved in the decision making process.

3. By increasing the standard of your information held in your database/data warehouse you are not just cleansing data, you are in fact adding a valuable commodity to the businesses portfolio. In today's business world data is an extremely sort after product, many companies sell on information in regard to their business practice. Accurate data can increase the importance and rate of the data your company holds, giving additional 'value-add' to a data-cleansing project.

## **Summary**

[TOP](#)

As has been discussed, in our vast experience in data warehousing and data cleansing, companies frequently over look a fundamental component for success; clean and reliable data.

Millions of pounds/dollars are spent on hardware and software for storing, manipulating and reading data. By spending just a fraction of that budget on implementing business rules, data checks and general data cleansing initiative a wide variety of gains can be made to your overall business. Investing in data, its quality and business intelligence a guaranteed increase in business efficiency, customer relations and profitability can be made to your company.