Data Quality Tools for Data Warehousing – A Small Sample Survey

Using Information in Government Program

Center for Technology in Government
University at Albany / SUNY

Data Quality Tools for Data Warehousing - A Small Sample Survey

Abstract

It is estimated that as high as 75% of the effort spent on building a data warehouse can be attributed to back-end issues, such as readying the data and transporting it into the data warehouse (Atre, 1998). Data quality tools are used in data warehousing to ready the data and ensure that clean data populates the warehouse, thus enhancing usability of the warehouse. This research focuses on the problems in the data that are addressed by data quality tools. Specific questions of the data can elicit information that will determine which features of the data quality tools are appropriate in which circumstances. The objective of the effort is to develop a tool to support the identification of data quality issues and the selection of tools for addressing those issues. A secondary objective is to provide information on specific tools regarding price, platform, and unique features of the tool.

Introduction

Attention to data quality is a critical issue in all areas of information resources management. A recent article in the Wall Street Journal (7/13/98) relates the domino effect that occurs when erroneous information is typed into a central database. A new airport in Hong Kong suffered catastrophic problems in baggage handling, flight information, and cargo transfer. The ramifications of the dirty data were felt throughout the airport. Flights took off without luggage, airport officials tracked flights with plastic pieces on magnetic boards, and airlines called confused ground staff on cellular phones to let them know where even more confused passengers could find their planes (Arnold, 1998). The new airport had been depending on the central database to be accurate. When it wasn't, the airport paid the price in terms of customer satisfaction and trust.

Data warehousing is emerging as the cornerstone of an organization's information infrastructure. It is imperative that the issue of data quality be addressed if the data warehouse is to prove beneficial to an organization. Corporations, government agencies and not-for-profit groups are all inundated with enormous amounts of data. The desire to use this data as a resource for the organization has increased the move towards data warehouses. This information has the potential to be used by an organization to generate greater understanding of their customers, processes, and the organization itself.

There potential to increase the usefulness of data by combining it with other data sources is great. But, if the underlying data is not accurate, any relationships found in the data warehouse will be misleading. For example, most payroll systems require a social security number when setting up an employee file. If no number is available when the file is set up, an incorrect number may be used, such as 111-11-1111, in order to facilitate payroll processing. The intention is that the numbers would be changed when the correct

social security number is obtained. If the numbers are not changed, then some relationship may exist in the database, but the relationship would be misleading because the underlying data is inaccurate.

Data Flow

The steps for building a data warehouse or repository are well understood. The data flows from one or more source databases into an intermediate staging area, and finally into the data warehouse or repository (see Figure 1). At each stage there are data quality tools available to massage and transform the data, thus enhancing the usability of the data once it resides in the data warehouse.
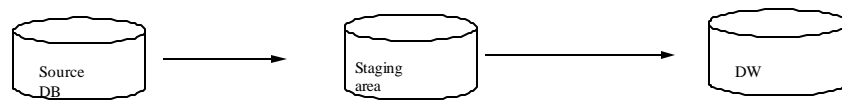
```
Source          Staging          DW
DB              area
```

Figure 1- Data Flow

Research Questions

This study was designed to address the following questions about the relationship between data quality tools and the data destined for the data warehouse:

1. How can a data quality tool help to ensure that data is clean before it enters the data warehouse?
2. What properties of the data need to be addressed in order to ensure that the data is clean?
3. What are the features of a data quality tool that will clean the data?
4. Which data quality tools contain the features that are needed to ensure clean and accurate data?
5. How do the various data quality tools compare in terms of price, platform and "user-friendliness?"

The research project was also designed to use the resulting information as the foundation for the "Mapping Data Problems to Features of Data Quality Tools" matrix.

Review of Data Quality in Data Warehouses

Estimates as high as 75% of the effort spent on a data warehouse are attributed to back-end issues, such as readying the data and transporting it into the data warehouse. Data cleansing activities account for nearly half of that time (Atre, 1998). Hundreds of tools are available to automate portions of the tasks associated with auditing, cleansing, extracting, and loading data into data warehouses. Most of these tools fall into the data extracting and loading classification while only a small number would be considered auditing or cleansing tools. Historically, IT personnel have developed their own routines for cleansing data. For example, data is validated on data entry based on what type of data should be in the field, reasonableness checks, and other validation checks. Data quality tools are emerging as a way to correct and clean data at many stages in building and maintaining a data warehouse. These tools are used to audit the data at the source, transform the data so that it is consistent throughout the warehouse, segment the data into atomic units, and ensure the data matches the business rules. The tools can be stand-alone packages, or can be integrated with data warehouse packages.

Data flows from the source database into an intermediate staging area, and then into a data warehouse. Different tools can be used at each stage. (see Figure 2)
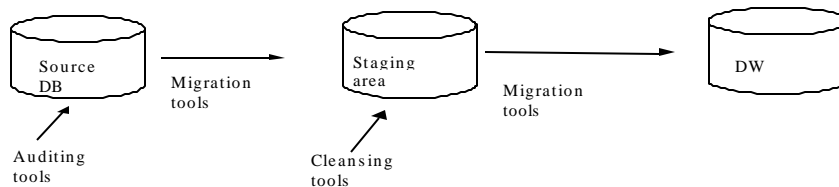

Figure 2- Data Flow and Data Quality Tools

Data Quality Tools

Data quality tools generally fall into one of three categories: auditing, cleansing and migration. The focus of this paper is on tools that clean and audit data, with a limited look at tools that extract and migrate data.

Data auditing tools enhance the accuracy and correctness of the data at the source. These tools generally compare the data in the source database to a set of business rules. (Williams, 1997) When using a source external to the organization, business rules can be determined by using data mining techniques to uncover patterns in the data. Business rules that are internal to the organization should be entered in the early stages of

evaluating data sources. Lexical analysis may be used to discover the business sense of words within the data. The data that does not adhere to the business rules could then be modified as necessary.

Data cleansing tools are used in the intermediate staging area. The tools in this category have been around for a number of years. A data cleansing tool cleans names, addresses and other data that can be compared to an independent source. These tools are responsible for parsing, standardizing, and verifying data against known lists such as U.S. Postal Codes. The data cleansing tools contain features which perform the following functions:

- Data parsing (elementizing)- breaks a record into atomic units that can be used in subsequent steps. Parsing includes placing elements of a record into the correct fields. In the following example "ST" is used in a variety of ways:
  Elizabeth St. Francis
  1130 1st St.
  St. 101
  St. Paul, MN 50505
- Data standardization- converts the data elements to forms that are standard throughout the data warehouse. For example, all incidences of avenue should be represented as ave., not Avenue, avenue, or av.
- Data correction and verification- matches data against know lists, such as U.S. Postal Codes, product lists, internal customer lists.
- Record matching- determines whether two records represent data on the same subject. For example, the following two records probably represent the same person:
Sue Smith		Suzanne Smith
19 Rt 9G	AND	19 North Road
Hyde Park, NY 12538		Hyde Park, NY 12538
(914)229-1111		(914)229-1111
- Data transformation- ensures consistent mapping between source systems and data warehouse. For example, "1" for male, and "2" for female becomes "M" and "F".
- Householding – combining individual records that have the same address.
- Documenting – documenting the results of the data cleansing steps in the meta data

The third type of tool, the data migration tool, is used in extracting data from a source database, and migrating the data into an intermediate storage area. The migration tools also transfer data from the staging area into the data warehouse. The data migration tool is responsible for converting the data from one platform to another. A migration tool will map the data from the source to the data warehouse. It can also check for Y2K compliance and other simple cleansing activities. There can be a great deal of overlap in these tools and many of the same features are found in tools of each category.

Methodology

A review of the literature indicates that there are hundreds of tools that can be classified as data extraction, loading and cleansing tools. Only a small percentage of the tools perform the data cleansing and auditing functions, the majority perform extraction and loading functions. The majority of the auditing and cleansing tools were evaluated for this paper, as well as a small sample of the extraction and loading tools. Phase I of the literature review generated a preliminary list of questions for the data. Phase II refined the list and used it as the foundation for building a matrix to be used in evaluating data sources.

The original plan was to obtain copies of the software tools for testing. This proved to be very difficult. Only one vendor had a demo version available. The other vendors had marketing demos, which showed screen shots of the software, but did not allow testing the software with live data. The opportunity to talk to the vendors, and see the software, became available at the DCI Data Warehouse Conference in New York City. The next step was to talk to the vendors to gather information related to price, platform, and additional features of the tools. I attended the conference on July 30, 1998 and interviewed 8 software vendors. The interviews consisted of questions of the vendors and demonstrations of the software.

The matrices were developed with the information gathered during the literature review, the product review and the interviews.

Results

The objective of the effort is to develop a tool to support the identification of data quality issues and the selection of tools for addressing those issues. In order to determine what features would be needed, the following questions were initially developed to be asked of the data:

- Is your data complete and valid?
- Does your data contain fields that must be split into smaller parts before entering the data warehouse?
- Does your data have abbreviations that should be changed to insure consistency throughout the data warehouse?
- Is your data correct?
- Is there redundancy in your data (is the same information in more than one place in the various databases that you are drawing from)?
- Do different forms of data need to be converted to a single form for consistency across the data warehouse?
- How well does the data reflect the business rules? Do you have missing values, illegal values, inconsistent values, or invalid relationships?
- Do you have free form text that needs to be indexed and classified to be useful in the data warehouse?

This list of questions was reviewed by members of four New York State agencies in the initial stages of developing the framework for data repositories within the agencies. Based on their review, additional questions were added.

A matrix (Table 1) was developed that mapped the features of the data quality tools to the questions that were asked. Examples of tools that contain the features are also part of the matrix. The matrix was reviewed by IT professionals from four New York State agencies. Based on their review, additional questions were added to the matrix. This matrix can be used by builders of data warehouses in the initial stages of development to evaluate their data sources. Once the questions have been asked of the data, the warehouse developer will be able to identify problems in the data sources. The data quality tools have different features to address specific problems in the data. The "Mapping Data Problems to Features of Data Quality Tools" matrix in Table 1 will allow the warehouse developer to focus on which features are needed to address specific problems in the data sources. For example, if the data sources contain primarily name and address data, then a data cleansing tool may be sufficient. On the other hand, if most of the data is financial, then an auditing tool may be more appropriate.

Table 2 contains information about specific tools, including URL's, price, platform, and special features of the tool. The matrix can be used to begin evaluation of specific tools.

Table 1- Mapping Data Problems to Features of Data Quality Tools

Questions to be asked	Features	Tools
Auditing Tools		
Is your data complete and valid?	Data examination- determines quality of data, patterns within it, and number of different fields used	WizSoft- WizRule Vality- Integrity
Does your data comply to your business rules? (Do you have missing values, illegal values, inconsistent values, invalid relationships?)	Compare to business rules and assess data for consistency and completeness against rules	Prism Solutions, Inc.- Prism Quality Manager WizSoft - WizRule Vality- Integrity
Are you using sources that comply to your business rules?	Data reengineering- examining the data to determine what the business rules are	WizSoft – WizRule Vality- Integrity
Cleansing Tools		
Does your data need to be broken up between source and data warehouse?	Data parsing (elementizing)- context and destination of each component of each field	Trillium Software- Parser i.d. Centric- DataRight
Does your data have abbreviations that should be changed to insure consistency?	Data standardizing- converting data elements to forms that are standard throughout the DW	Trillium Software- Parser i.d. Centric- DataRight
Is your data correct?	Data correction and verification- matches data against known lists (addresses, product lists, customer lists)	Trillium Software- Parser Trillium Software- GeoCoder i.d. Centric- ACE, Clear I.D. Library Group 1- NADIS
Is there redundancy in your data?	Record matching- determines whether two records represent data on the same object	Trillium Software- Matcher Innovative Systems- Match i.d. Centric- Match/Consolidation Group 1- Merge/Purge Plus
Are there multiple versions of company names in your database?	Record matching- based on user specified fields such as tax ID	Innovative Systems- Corp-Match
Is your data consistent prior to entering data warehouse?	Transform data- "1" for male, "2" for female becomes "M" & "F"- ensures consistent mapping between source systems and data warehouse	Vality- Integrity i.d. Centric- Match/Consolidation
Do you have information in free form fields that differs between databases?	Data reengineering- examining the data to determine what the business rules are	Vality- Integrity
Do you multiple individuals in the same household that need to be grouped together?	Householding- combining individual records that have same address	i.d. Centric- Match/Consolidation Trillium Software- Matcher
Does your data contain atypical words- such as industry specific words, ethnic or hyphenated names?	Data parsing combined with data verification- comparison to industry specific lists	i.d. Centric- ACE, Clear I.D.
Migration and Other Tools		
Do you have multiple formats to be accessed- relational dbs, flat files, etc.?	Access the data then map it to the dw schema	Enterprise/Integrator by Carleton.
Do you have free form text that needs to be indexed, classified, other?	Text mining- extracts meaning and relevance from large amounts of information	Semio- SemioMap
Have the rules established during the data cleansing steps been reflected in the metadata?	Documenting- documenting the results of the data cleansing steps in the metadata	Vality- Integrity
Is data Y2K compliant?	Data verifiacation within a migration tool	Enterprise/Integrator by Carleton.

Table 2 - Data Quality Products

Product	Company	URL	Purpose	Price	Platform	Train-ing	Comments
ACE, Clear I.D., Data Right, match/consolidation	i.d. Centric	http://www.idcentric.com/products.html	Address Cleansing	$35-500K	UNIX, NT/95, AS400		Name and address cleansing, some enhancement by adding geocode and demographic information
Corp-Match	Innovative Systems	http://www.innovativesystems.net	Address Cleansing				
Code 1 Plus	Group 1 Software	http://www.g1.com/products/ds.asp?DS_ID=C1P	Address Cleansing	$10-50K	MVS/UNIX		Address cleaning only, no name changes
Merge/Purge Plus	Group 1 Software	http://www.g1.com/products/ds.asp?DS_ID=MPP	Address Cleansing	$10-20K	MVS/UNIX		Address cleaning only, no name changes
NADIS	Group 1	http://www.g1.com/products/ds.asp?DS_ID=NADIS	Address Cleansing	$125K	MVS/UNIX		Name and address cleansing-includes householding
WizRule	WizSoft, Inc.	http://www.wizsoft.com/	Audit	$1.4K	Windows		Audits data with data mining techniques, generates business rules from data and exceptions to rules
Enterprise/Integrator	Carleton	http://www.apertus.com/products/ei/index.htm	Audit/ Cleanse	$90K	ODBC	3 days included with purchase	Matching, consolidation, "smart updating" (reads target as source, dynamically updates, deletes, inserts, based on business rules)
Integrity	Vality	http://www.vality.com/	Audit/ Cleanse	$195-245K	UNIX/NT Mainframe		No built-in functions, everything

Conclusion

Data quality tools are available to enhance the quality of the data at several stages in the process of developing a data warehouse. Cleansing tools can be useful in automating many of the activities that are involved in cleansing the data- parsing, standardizing, correction, matching, transformation and householding. Many of the tools specialize in auditing the data, detecting patterns in the data, and comparing the data to business rules. Data extraction and loading tools are available to translate the data from one platform to another, and populate the data warehouse.

In the initial stages of data warehouse development the sources of the data should be examined. Questions should be asked of the data source that would enable the developer of the warehouse to know what problems exist with the data. Once these problems have been isolated, the warehouse builder could determine which features of the data quality tools address the specific needs of the data sources to be used. The matrix that has been developed will guide the warehouse developer towards the tool that would be appropriate for the data sources that will eventually populate the warehouse. Once the proper tools have been identified, the second matrix could be used compare price, platform, and special features of each tool. The two matrices work together to enable the data warehouse developer to efficiently choose the software tool suitable to the data sources that are to be used in the warehouse.

Further Research

Further research in this area is needed to validate the claims of the vendors. A sample database should be run against each of these tools and the cleanliness of the data should be verified after each iteration.

Selected References

Aragon, L. (1998). "Down With Dirt", PCWeek, February 9, 1998

Arnold, W. (1998). "Human Error Causes System Glitches the Embarrass New Asisan Airports", Wall Street Journal Interactive Edition. http://interactive.wsj.com.

Atre, S. (1998). "Rules for Data Cleansing", ComputerWorld.

English, L. (1998). "Data Quality: Meeting Customer Needs", Pitney Bowes white paper

English, L., (1996). "Help for Data Quality Problems", InformationWeek, October 7, 1996, pp. 53

Greenfield, L. (1998). "Data Cleaning, Extraction and Loading Tools" http://pwp.starnetinc.com/larryg/clean.html.

Haggerty, N. (1998). "Toxic Data", DM Review Magazine, June 1998

Horowitz, A. (1998). "Ensuring the Integrity of Your Data", Beyond Computing, May 1998

i.d. Centric (1998). "Customer Data Quality: Building the Foundation for a One-to-One Customer Relationship", i.d. Centric white paper

Kimball, R. (1996). "Dealing with Dirty Data" DBMS Online.
Moss, L. (1998). "Data Cleansing: A Dichotomy of Data Warehousing?", DM Review Magazine, February 1998

O'Neill, P. (1998). "It's a Dirty Job: Cleaning Data in the Warehouse", Gartner Group, January 12, 1998

Strange, K. (1997). "A Taxonomy of Data Quality", Gartner Group, May 29, 1997

Watterson, K. (1998). "Dirty Data, Dumb Decisions", DM Review Magazine, March 1998

Williams, J. (1997). "Tools for Traveling Data" DBMS Online, June 1997