

Pam Neely, Graduate Assistant, Center for Technology in Government

Pam Neely gave a presentation on how data problems can be mapped to features of data quality tools. She began by explaining the basic process of building a data warehouse. Data usually comes from one or several source databases, migrates to an intermediate storage area where the data is transformed, then it migrates to the data warehouse. Along the way, there are tools— auditing tools, cleansing tools and migration tools— to help ensure that the data is clean and accurate prior to being used to populate the data warehouse:

Auditing tools are used at the source, cleansing tools are used in the intermediate stage, and migration tools are used to move the data from the source to the staging area and then to the data warehouse. The data auditing tools ensure accuracy and correctness at the source of the data; they also compare the data to a set of business rules to perform validation checks. Data cleansing tools are used in the intermediate stage. They cleanse data by breaking the records into atomic units, standardizing the elements, matching records against each other to check for duplication, and consolidating data. Finally, data migration tools extract data from a source database, move it to the intermediate staging area, map data to the data warehouse schema, and move it into the warehouse. Pam presented a matrix mapping data quality problems to a selected set of data quality tools.

Table 2: Data Quality Matrix

Data Quality Tools

| Problems | 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-WizRuleVality-Integrity |
| How well does the data reflect the 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.-PrismQuality ManagerWizSoft-WizRuleVality-Integrity |
| Are you using sources that do not comply to your business rules? | Data reengineering-examining the data to determine what the business rules are | WizSoft-WizRuleVality-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-Parseri.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-Parseri.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-Integrityi.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. |

Data Quality Tools

| | | |
|--|---|----------------|
| 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 | |
| Is data Y2K compliant? | | |
| Is the quality of the data poor and people don't care because they have adjusted to it? | | |

Finally, Pam mentioned a few important questions that data quality tools cannot address:

- (1) Have the users of the source database adjusted to poor quality of the database and developed "work-arounds"?
- (2) Is there conflicting information that can't be compared to a known resource?
- (3) Do you have a lot of soft data that needs to be placed in your data warehouse?