

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 19 Rt 9G Hyde Park, NY 12538 (914)229-1111 AND Suzanne Smith 19 North Road Hyde Park, NY 12538 (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.