

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