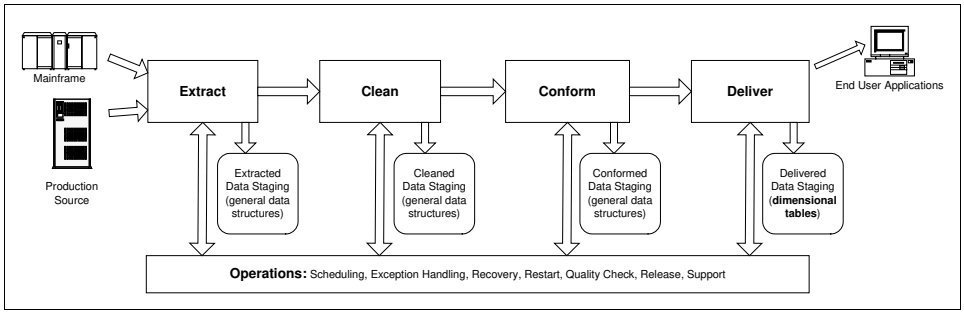

Figure 1.2 The Four Staging Steps of a Data Warehouse.

the data being transferred to the next stage. The central chapters of this book are organized around these steps. The four steps are:

1. **Extracting**. The raw data coming from the source systems is usually written directly to disk with some minimal restructuring but before significant content transformation takes place. Data from structured source systems (such as IMS databases, or XML data sets) often is written to flat files or relational tables in this step. This allows the original extract to be as simple and as fast as possible and allows greater flexibility to restart the extract if there is an interruption. Initially captured data can then be read multiple times as necessary to support the succeeding steps. In some cases, initially captured data is discarded after the cleaning step is completed, and in other cases data is kept as a long-term archival backup. The initially captured data may also be saved for at least one capture cycle so that the differences between successive extracts can be computed.

   
   We save the serious content transformations for the cleaning and conforming steps, but the best place to resolve certain legacy data format issues is in the extract step. These format issues include resolving repeating groups, REDEFINEs, and overloaded columns and performing low-level data conversions, including converting bit encoding to character, EBCDIC to ASCII, and packed decimal to integer. We discuss these steps in detail in Chapter 3.

2. **Cleaning.** In most cases, the level of data quality acceptable for the source systems is different from the quality required by the data warehouse. Data quality processing may involve many discrete steps, including checking for valid values (is the zip code present and is it in the range of valid values?), ensuring consistency across values (are the zip code and the city consistent?), removing duplicates (does the same customer appear twice with slightly different attributes?), and checking whether complex business rules and procedures have been enforced (does the Platinum customer have the associated extended