- LION Bioscience—provider of iDEA, a modular ADME predictive system. The absorption module predicts Caco-2 cell permeation, and performs dose-response modeling of the oral absorption. The metabolism module predicts first-pass effects and models metabolic parameters. Future modules are planned for distribution and elimination (124).
- PASS—prediction of biological activity spectra—compares a test structure with those in a database of about 45,000 structures with known activity/toxicity, using topological descriptors and probability calculations (125).

4.4 Property Calculations Online

Many of the providers of software and databases of chemical properties also provide online calculation services. These include Daylight Chemical Information (126), ACD labs (127), and Syracuse Research (128). In addition, the following sites provide online calculations of a variety of properties:

- Molinspiration—calculates Log*P*, polar surface area, Lipinski Rule-of-5, and a drug-likeness index (129)
- Alogp—VCCLab online Log*P* calculation (130)
- PETRA—The University of Erlangen property calculation routines (131)
- a USEPA Suite—including implementations of the Syracuse Research software (132)

5 DATA WAREHOUSES AND DATA MARTS

Even relational databases have their limitations when dealing with huge amounts of data and high user traffic. The burden of continual data updating and registration—activities known as OnLine Transaction Processing (OLTP), can considerably slow down searching and report generation activity—known as OnLine Analytical Processing (OLAP). For this reason, it is becoming common in the database field to build special large databases designed primarily for searching purposes—so-called data warehouses (133). These warehouses have pre-computed indexes and tables that facilitate repeated searching. A special database architecture, known as the star schema, facilitates OLAP activity. In this design, one or more large fact tables contain records of frequently searched data for each object (e.g., structure or reaction) in the database. The fact table is joined to smaller dimension tables that contain the relational information. The schema is known as a star schema because the architecture resembles a many-pointed star, with the fact table at the center, and dimension tables at the ends of the arms. The design of the fact and dimension tables in the warehouse should reflect the searching habits of the users to get the best performance. Probably the first mention of data warehousing in the pharmaceutical area was that of Axel and Song in 1997 (134).

5.1 Data Warehouses of Chemical Information

A data warehouse is designed to consolidate structures and data from many diverse sources, including relational databases, flat databases, and structure and data files. It is considered to be multidimensional. A true chemical data warehouse might contain sequences, 2D structures, 3D models, Markush structures, and reactions—all in the same database. No such commercial database currently exists, but databases presently being developed at MDL and other vendors are examples of chemical data warehouses of structures and their reactions. The MDL data warehouse framework is termed the concordance. The fact table for the concordance is the source table, which brings together structure and reaction identifiers from all the various data sources and links them to the unique structures in the warehouse (Fig. 9.16). Using the concordance, a substructure search can retrieve a set of unique, unduplicated structures, along with pointers to all the relevant identifiers and reactions in the various data sources. Similar pointers exist to the original citations and stored data.

Physicochemical properties that are based solely on the structure are stored in the data warehouse, but properties that are data-source dependent, such as citation or biological activity, are only referenced. A typical use of a chemical warehouse is to search for a set of