"*How does data mining relate to information processing and on-line analytical processing?*"

Information processing, based on queries, can find useful information. However, answers to such queries reflect the information directly stored in databases or computable by aggregate functions. They do not reflect sophisticated patterns or regularities buried in the database. Therefore, information processing is not data mining.

On-line analytical processing comes a step closer to data mining since it can derive information summarized at multiple granularities from user-specified subsets of a data warehouse. Such descriptions are equivalent to the class/concept descriptions discussed in Chapter 1. Since data mining systems can also mine generalized class/concept descriptions, this raises some interesting questions: Do OLAP systems perform data mining? Are OLAP systems actually data mining systems?

The functionalities of OLAP and data mining can be viewed as disjoint: OLAP is a data summarization/aggregation *tool* which helps simplify data analysis, while data mining allows the *automated discovery* of implicit patterns and interesting knowledge hidden in large amounts of data. OLAP tools are targeted toward simplifying and supporting interactive data analysis, but the goal of data mining tools is to automate as much of the process as possible, while still allowing users to guide the process. In this sense, data mining goes one step beyond traditional on-line analytical processing.

An alternative and broader view of data mining may be adopted in which data mining covers both data description and data modeling. Since OLAP systems can present general descriptions of data from data warehouses, OLAP functions are essentially for user-directed data summary and comparison (by drilling, pivoting, slicing, dicing, and other operations). These are, though limited, data mining functionalities. Yet according to this view, data mining covers a much broader spectrum than simple OLAP operations because it not only performs data summary and comparison, but also performs association, classification, prediction, clustering, time-series analysis, and other data analysis tasks.

Data mining is not confined to the analysis of data stored in data warehouses. It may analyze data existing at more detailed granularities than the summarized data provided in a data warehouse. It may also analyze transactional, textual, spatial, and multimedia data which are difficult to model with current multidimensional database technology. In this context, data mining covers a broader spectrum than OLAP with respect to data mining functionality and the complexity of the data handled.

Since data mining involves more automated and deeper analysis than OLAP, data mining is expected to have broader applications. Data mining can help business managers find and reach more suitable customers, as well as gain critical business insights that may help to drive market share and raise profits. In addition, data mining can help managers understand customer group characteristics and develop optimal pricing strategies accordingly, correct item bundling based not on intuition but on actual item groups derived from customer purchase patterns, reduce promotional spending and at the same time, increase net effectiveness of promotions overall.

2.6.2 From on-line analytical processing to on-line analytical mining

In the field of data mining, substantial research has been performed for data mining at various platforms, including transaction databases, relational databases, spatial databases, text databases, time-series databases, flat files, data warehouses, etc.

Among many different paradigms and architectures of data mining systems, **On-Line Analytical Mining (OLAM)** (also called **OLAP mining**), which integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multidimensional databases, is particularly important for the following reasons.

1. **High quality of data in data warehouses.** Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data transformation, and data integration as preprocessing steps. A data warehouse constructed by such preprocessing serves as a valuable source of high quality data for OLAP as well as for data mining. Notice that data mining may also serve as a valuable tool for data cleaning and data integration as well.

2. **Available information processing infrastructure surrounding data warehouses.** Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple, heterogeneous databases, ODBC/OLEDB connections, Web-accessing and service facilities, reporting and OLAP

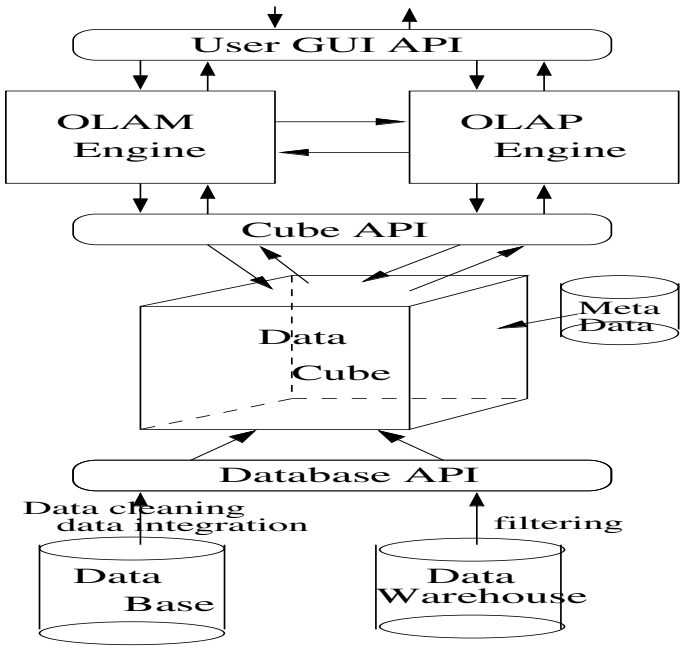

Figure 2.21: An integrated OLAM and OLAP architecture.

analysis tools. It is prudent to make the best use of the available infrastructures rather than constructing everything from scratch.

3. **OLAP-based exploratory data analysis.** Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms. On-line analytical mining provides facilities for data mining on different subsets of data and at different levels of abstraction, by drilling, pivoting, filtering, dicing and slicing on a data cube and on some intermediate data mining results. This, together with data/knowledge visualization tools, will greatly enhance the power and flexibility of exploratory data mining.

4. **On-line selection of data mining functions.** Often a user may not know what kinds of knowledge that she wants to mine. By integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

Architecture for on-line analytical mining

An OLAM engine performs analytical mining in data cubes in a similar manner as an OLAP engine performs on-line analytical processing. An integrated OLAM and OLAP architecture is shown in Figure 2.21, where the OLAM and OLAP engines both accept users' on-line queries (or commands) via a User_GUI_API and work with the data cube in the data analysis via a Cube_API. A metadata directory is used to guide the access of the data cube. The data cube can be constructed by accessing and/or integrating multiple databases and/or by filtering a data warehouse via a Database_API which may support OLEDB or ODBC connections. Since an OLAM engine may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, etc., it usually consists of multiple, integrated data mining modules and is more sophisticated than an OLAP engine.

The following chapters of this book are devoted to the study of data mining techniques. As we have seen, the introduction to data warehousing and OLAP technology presented in this chapter is essential to our study of data mining. This is because data warehousing provides users with large amounts of clean, organized, and summarized data, which greatly facilitates data mining. For example, rather than storing the details of each sales transaction, a data warehouse may store a summary of the transactions per item type for each branch, or, summarized to a higher level, for each country. The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining.

Moreover, we also believe that data mining should be a human-centered process. Rather than asking a data mining system to generate patterns and knowledge automatically, a user will often need to interact with the system to perform exploratory data analysis. OLAP sets a good example for interactive data analysis, and provides the necessary preparations for exploratory data mining. Consider the discovery of association patterns, for example. Instead of mining associations at a primitive (i.e., low) data level among transactions, users should be allowed to specify roll-up operations along any dimension. For example, a user may like to roll-up on the *item* dimension to go from viewing the data for particular TV sets that were purchased to viewing the brands of these TVs, such as SONY or Panasonic. Users may also navigate from the transaction level to the customer level or customer-type level in the search for interesting associations. Such an OLAP-style of data mining is characteristic of OLAP mining.

In our study of the principles of data mining in the following chapters, we place particular emphasis on OLAP mining, that is, on the *integration of data mining and OLAP technology.*

2.7 Summary

- **A data warehouse** is a *subject-oriented, integrated, time-variant*, and *nonvolatile* collection of data organized in support of management decision making. Several factors distinguish data warehouses from operational databases. Since the two systems provide quite different functionalities and require different kinds of data, it is necessary to maintain data warehouses separately from operational databases.
- **A multidimensional data model** is typically used for the design of corporate *data warehouses* and *departmental data marts*. Such a model can adopt either a *star schema*, *snowflake schema*, or *fact constellation schema*. The core of the *multidimensional model* is the **data cube**, which consists of a large set of *facts* (or *measures*) and a number of *dimensions*. Dimensions are the entities or perspectives with respect to which an organization wants to keep records, and are hierarchical in nature.
- **Concept hierarchies** organize the values of attributes or dimensions into gradual levels of abstraction. They are useful in mining at multiple levels of abstraction.
- **On-line analytical processing (OLAP)** can be performed in data warehouses/marts using the multidimensional data model. Typical OLAP operations include *roll-up*, *drill-*(*down, cross, through*), *slice-and-dice*, *pivot* (*rotate*), as well as statistical operations such as ranking, computing moving averages and growth rates, etc. OLAP operations can be implemented efficiently using the data cube structure.
- Data warehouses often adopt a **three-tier architecture**. The bottom tier is a *warehouse database server*, which is typically a relational database system. The middle tier is an *OLAP server*, and the top tier is a *client*, containing query and reporting tools.
- OLAP servers may use **Relational OLAP (ROLAP)**, or **Multidimensional OLAP (MOLAP)**, or **Hybrid OLAP (HOLAP)**. A ROLAP server uses an extended relational DBMS that maps OLAP operations on multidimensional data to standard relational operations. A MOLAP server maps multidimensional data views directly to array structures. A HOLAP server combines ROLAP and MOLAP. For example, it may use ROLAP for historical data while maintaining frequently accessed data in a separate MOLAP store.
- A data cube consists of a **lattice of cuboids**, each corresponding to a different degree of summarization of the given multidimensional data. **Partial materialization** refers to the selective computation of a subset of the cuboids in the lattice. **Full materialization** refers to the computation of all of the cuboids in the lattice. If the cubes are implemented using MOLAP, then **multiway array aggregation** can be used. This technique "overlaps" some of the aggregation computation so that full materialization can be computed efficiently.
- OLAP query processing can be made more efficient with the use of indexing techniques. In **bitmap indexing**, each attribute has its own bimap vector. Bitmap indexing reduces join, aggregation, and comparison operations to bit arithmetic. **Join indexing** registers the joinable rows of two or more relations from a relational database, reducing the overall cost of OLAP join operations. **Bitmapped join indexing**, which combines the bitmap and join methods, can be used to further speed up OLAP query processing.
- Data warehouse **metadata** are data defining the warehouse objects. A metadata repository provides details regarding the warehouse structure, data history, the algorithms used for summarization, mappings from the source data to warehouse form, system performance, and business terms and issues.

- A data warehouse contains **back-end tools and utilities** for populating and refreshing the warehouse. These cover data extraction, data cleaning, data transformation, loading, refreshing, and warehouse management.
- **Discovery-driven exploration** of data cubes uses precomputed measures and visual cues to indicate data exceptions, guiding the user in the data analysis process, at all levels of aggregation. **Multifeature cubes** compute complex queries involving multiple dependent aggregates at multiple granularities. The computation of cubes for discovery-driven exploration and of multifeature cubes can be achieved efficiently by taking advantage of efficient algorithms for standard data cube computation.
- Data warehouses are used for *information processing* (querying and reporting), *analytical processing* (which allows users to navigate through summarized and detailed data by OLAP operations), and *data mining* (which supports knowledge discovery). OLAP-based data mining is referred to as **OLAP mining**, or on-line analytical mining (**OLAM**), which emphasizes the interactive and exploratory nature of OLAP mining.

Exercises

1. State why, for the integration of multiple, heterogeneous information sources, many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach.
2. Design a data warehouse for a regional weather bureau. The weather bureau has about 1,000 probes which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. All data are sent to the central station, which has collected such data for over 10 years. Your design should facilitate efficient querying and on-line analytical processing, and derive general weather patterns in multidimensional space.
3. What are the differences between the three typical methods for modeling multidimensional data: the *star* model, the *snowflake* model, and the *fact constellation* model? What is the difference between the *star* warehouse model and the *starnet* query model? Use an example to explain your point(s).
4. A popular data warehouse implementation is to construct a multidimensional database, known as a data cube. Unfortunately, this may often generate a huge, yet very sparse multidimensional matrix.
   (a) Present an example, illustrating such a huge and sparse data cube.
   (b) Design an implementation method which can elegantly overcome this sparse matrix problem. Note that you need to explain your data structures in detail and discuss the space needed, as well as how to retrieve data from your structures, and how to handle incremental data updates.
5. Data warehouse design:
   (a) Enumerate three classes of schemas that are popularly used for modeling data warehouses.
   (b) Draw a schema diagram for a data warehouse which consists of three dimensions: *time, doctor*, and *patient*, and two measures: *count*, and *charge*, where *charge* is the fee that a doctor charges a patient for a visit.
   (c) Starting with the base cuboid $(day, doctor, patient)$, what specific OLAP operations should be performed in order to list the total fee collected by each doctor in VGH (Vancouver General Hospital) in 1997?
   (d) To obtain the same list, write an SQL query assuming the data is stored in a relational database with the schema.
   $$fee(day, month, year, doctor, hospital, patient, count, charge).$$
6. Computing measures in a data cube:
   (a) Enumerate three categories of measures, based on the kind of aggregate functions used in computing a data cube.
   (b) For a data cube with three dimensions: *time, location*, and *product*, which category does the function *variance* belong to? Describe how to compute it if the cube is partitioned into many chunks.
   Hint: The formula for computing *variance* is: $\frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \bar{x_i}^2$, where $\bar{x_i}$ is the average of x_i's.