DBMS AND DATAMINING

Mr. Amit Taneja (Assistant Professor) , Mukti Vashishtha

Department of Computer Applications
Master of Computer Applications
Invertis University, Bareilly

Abstract:

A **database** is an organized collection of data, generally stored and accessed electronically from a computer system. Where databases are more complex they are often developed using formal design and modeling techniques.

The database management system (DBMS) is the software that interacts with end users, applications, and the database itself to capture and analyze the data. The DBMS software additionally encompasses the core facilities provided to administer the database. The sum total of the database, the DBMS and the associated applications can be referred to as a "database system". Often the term "database" is also used to loosely refer to any of the DBMS, the database system or an application associated with the database.

Computer scientists may classify database-management systems according to the database models that they support. Relational databases became dominant in the 1980s. These model data as rows and columns in a series of tables, and the vast majority use SQL for writing and querying data. In the 2000s, non-relational databases became popular, referred to as NoSQL because they use different query languages.

Terminology and Overview

Formally, a "database" refers to a set of related data and the way it is organized. Access to this data is usually provided by a "database management system" (DBMS) consisting of an integrated set of computer software that allows users to interact with one or more databases and provides access to all of the data contained in the database (although restrictions may exist that limit access to particular data). The DBMS provides various functions that allow entry, storage and retrieval of large quantities of information and provides ways to manage how that information is organized.

Because of the close relationship between them, the term "database" is often used casually to refer to both a database and the DBMS used to manipulate it.

Outside the world of professional information technology, the term *database* is often used to refer to any collection of related data (such as a spreadsheet or a card index) as size and usage requirements typically necessitate use of a database management system.[1]

Existing DBMSs provide various functions that allow management of a database and its data which can be classified into four main functional groups:

- **Data definition** – Creation, modification and removal of definitions that define the organization of the data.
- **Update** – Insertion, modification, and deletion of the actual data.[2]
- **Retrieval** – Providing information in a form directly usable or for further processing by other applications. The retrieved data may be made available in a form basically the same as it is stored in the database or in a new form obtained by altering or combining existing data from the database.[3]
- **Administration** – Registering and monitoring users, enforcing data security, monitoring performance, maintaining data integrity, dealing with concurrency control, and recovering information that has been corrupted by some event such as an unexpected system failure.[4]

Both a database and its DBMS conform to the principles of a particular database model.[5] "Database system" refers collectively to the database model, database management system, and database.[6]

Physically, database servers are dedicated computers that hold the actual databases and run only the DBMS and related software. Database servers are usually multiprocessor computers, with generous memory and RAID disk arrays used for stable storage. Hardware database accelerators, connected to one or more servers via a high-speed channel, are also used in large volume transaction processing environments. DBMSs are found at the heart of most database applications. DBMSs may be built around a custom multitasking kernel with built-in networking support, but modern DBMSs typically rely on a standard operating system to provide these functions.[*citation needed*]

Since DBMSs comprise a significant market, computer and storage vendors often take into account DBMS requirements in their own development plans.[7]

Databases and DBMSs can be categorized according to the database model(s) that they support (such as relational or XML), the type(s) of computer they run on (from a server cluster to a mobile phone), the query language(s) used to access the database (such as SQL or XQuery), and their internal engineering, which affects performance, scalability, resilience, and security.

History

The sizes, capabilities, and performance of databases and their respective DBMSs have grown in orders of magnitude. These performance increases were enabled by the technology progress in the areas of processors, computer memory, computer storage, and computer networks. The concept of a database was made possible by the emergence of direct access storage media such as magnetic disks, which became widely available in the mid 1960s; earlier systems relied on sequential storage of data on magnetic tape. The subsequent development of database technology can be divided into three eras based on data model or structure: navigational,[8] SQL/relational, and post-relational.

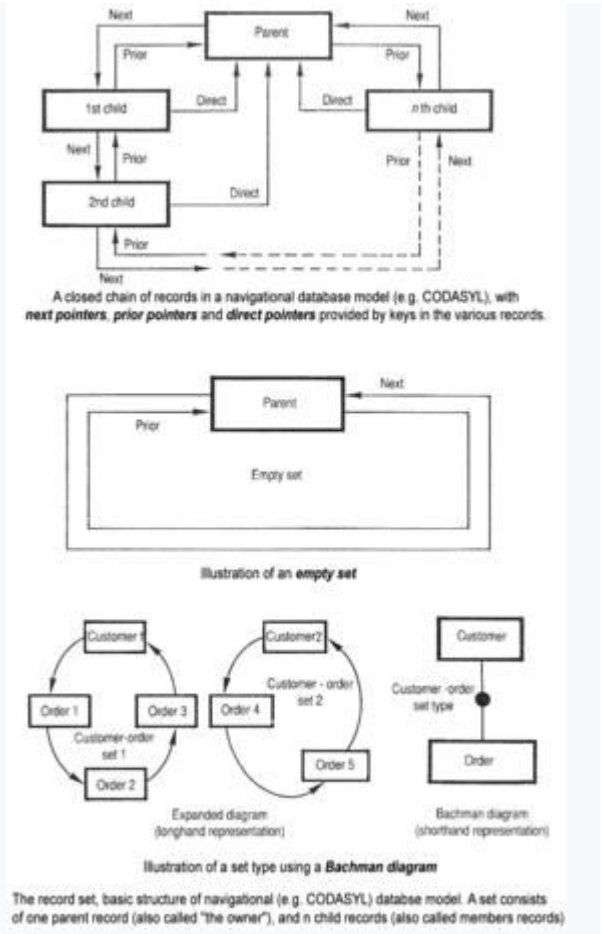
The two main early navigational data models were the hierarchical model and the CODASYL model (network model). These were characterized by the use of pointers (often physical disk addresses) to follow relationships from one record to another.

The relational model, first proposed in 1970 by Edgar F. Codd, departed from this tradition by insisting that applications should search for data by content, rather than by following links. The relational model employs sets of ledger-style tables, each used for a different type of entity. Only in the mid-1980s did computing hardware become powerful enough to allow the wide deployment of relational systems (DBMSs plus applications). By the early 1990s, however,

relational systems dominated in all large-scale data processing applications, and as of 2018 they remain dominant: IBM DB2, Oracle, MySQL, and Microsoft SQL Server are the most searched DBMS.[9] The dominant database language, standardised SQL for the relational model, has influenced database languages for other data models.[*citation needed*]

Object databases were developed in the 1980s to overcome the inconvenience of object–relational impedance mismatch, which led to the coining of the term "post-relational" and also the development of hybrid object–relational databases.

The next generation of post-relational databases in the late 2000s became known as NoSQL databases, introducing fast key–value stores and document-oriented databases. A competing "next generation" known as NewSQL databases attempted new implementations that retained the relational/SQL model while aiming to match the high performance of NoSQL compared to commercially available relational DBMSs.



Basic structure of navigational CODASYL database model

1960s, navigational DBMS

Further information: Navigational database

The introduction of the term *database* coincided with the availability of direct-access storage (disks and drums) from the mid-1960s onwards. The term represented a contrast with the tape-based systems of the past, allowing shared interactive use rather than daily batch processing. The Oxford English Dictionary cites a 1962 report by the System Development Corporation of California as the first to use the term "data-base" in a specific technical sense.[10]

As computers grew in speed and capability, a number of general-purpose database systems

emerged; by the mid-1960s a number of such systems had come into commercial use. Interest in a standard began to grow, and Charles Bachman, author of one such product, the Integrated Data Store (IDS), founded the Database Task Group within CODASYL, the group responsible for the creation and standardization of COBOL. In 1971, the Database Task Group delivered their standard, which generally became known as the *CODASYL approach*, and soon a number of commercial products based on this approach entered the market.

The CODASYL approach offered applications the ability to navigate around a linked data set which was formed into a large network. Applications could find records by one of three methods:

1. Use of a primary key (known as a CALC key, typically implemented by hashing)
2. Navigating relationships (called *sets*) from one record to another
3. Scanning all the records in a sequential order

Later systems added B-trees to provide alternate access paths. Many CODASYL databases also added a declarative query language for end users (as distinct from the navigational API). However CODASYL databases were complex and required significant training and effort to produce useful applications.

IBM also had their own DBMS in 1966, known as Information Management System (IMS). IMS was a development of software written for the Apollo program on the System/360. IMS was generally similar in concept to CODASYL, but used a strict hierarchy for its model of data navigation instead of CODASYL's network model. Both concepts later became known as navigational databases due to the way data was accessed: the term was popularized by Bachman's 1973 Turing Award presentation *The Programmer as Navigator*. IMS is classified by IBM as a hierarchical database. IDMS and Cincom Systems' TOTAL database are classified as network databases. IMS remains in use as of 2014.

The data mining context

The above data mining definition consists of three parts that must be properly qualified.

First, non-trivial discovery of relevant information implies the detection of patterns, tendencies and correlations that cannot be exposed through conventional query techniques, either because these are, in fact, inappropriate, or highly inefficient for the complexity of the problem. By contrast, data mining provides methods coming from disciplines such as artificial intelligence (machine learning) and multivariate analysis to address this kind of problems. These methods, based on statistically robust algorithms, can model complex relationships in structured and semi-structured data sets, involving different variable types, high scattering levels, and with no assumption about the underlying data distribution. Data mining modeling methods are usually

categorized as supervised learning (classification, regression, time series forecasting) and unsupervised learning (clustering, association rules detection, sequential patterns discovery).

Secondly, information discovery requires a methodology. It is needed in order to define the problem that must be addressed, the business context and the required analysis framework. The framework covers the variables or features that will be included in the analysis, as well as the sequence of preparation steps and modeling tasks to be conducted until a valid solution is found and can be applied (**Figure** 1). This methodology is translated into a set of processes that include the initial qualification and preparation of the data sets, the development, assembly and validation of one or several models for knowledge extraction and, finally, and most importantly, the deployment of the preparation-modeling stream into production and its continuous monitoring and recalibration. As in other environments, these processes are implemented using several underlying IT services that execute them.

Last but not least, the main goal of the extracted information is to improve business performance. First, exposing the new knowledge and insight in order to support strategic plans and tactical decisions. Then, deploying this knowledge into business operations by applying the analytical models and integrating the results inside the informational and transactional processes.

The business focus, the methodological approach and a service-oriented implementation, make data mining a core business analytics specialization, and not just a bunch of mathematical techniques and algorithms.

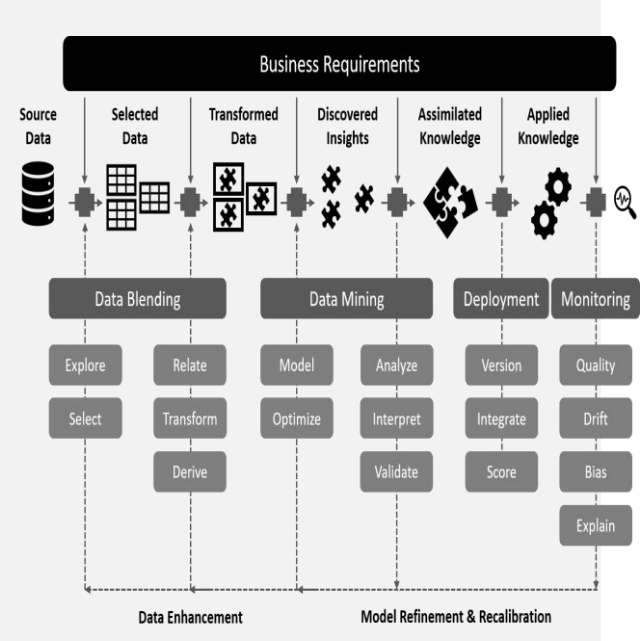


Figure 1 — Data mining solution workflow

Technological approaches for data mining: a historical review

In the last decade, the number of application areas and potential users of data mining has expanded considerably. Something that in the past was considered a restricted domain of highly skilled statistical practitioners now has evolved to be part of business applications, involving data engineers, developers and end-users. This step has been possible to the extent the technological evolution has facilitated and automated the use and application of the modeling techniques. The formalization and standardization of the methodology have also contributed, but not as decisively as the technology has enhanced a new generation of analytical applications where data mining is embedded, enabling business users to solve complex problems and take advantage of new opportunities.

Conclusions

There should be an additional sixth reason to finalize this exposition, and this is quite simple: the relational database system is a valuable resource in the analyst toolbox to build and integrate data mining operations. And there are many business scenarios where this option should always be considered. It is a matter of bringing the algorithm where corporate data resides.

In present times, where there is some obsession in renaming and reinventing every technology and discipline with more than five years, someone will soon say this is Edge Computing. We will see.