Mathematical Foundations of Database Management Systems and Modern Development Trends in the Field

Giorgi Ghlonti[1], Maguli Papiashvili[2]

[1]Muskhelishvili Institute of Computational Mathematics. ghlonti.g@gtu.ge

[2]Muskhelishvili Institute of Computational Mathematics. papiashvili.m@gtu.ge

Abstract

Some aspects of the organization of databases and database management systems, history of their development from the first network and hierarchical databases to modern online distributed systems are discussed.

It is shown how the relational approach was developed in response to the difficulties associated with the use of hierarchical and network schemas in database management systems.

The works related to the design and implementation of information systems, held from the 1960s in Muskhelishvili Institute of Computational Mathematics are described.

The concept of building large-scale, pervasive, distributed data processing system for accumulation and management of analytical information resources is established. It refers to design of cyber-infrastructure for support of the formation of analytical information resources throughout the country and its delivery to end-user.

Keywords: *Database Management System; Relational model; Distributed databases; Pervasive cyber-infrastructure.*

A database represents a logically interconnected information, united by a certain sign, to which a formalized description of its structure (the so-called conceptual schema of the database) is attached.

This allows to apply as specialized package – a database management system (DBMS) – to access the data, and to turn the data into a public information resource.

The conceptual schema, in the simplest case, represented a set of table structures, where for each record of the database information about mutual arrangement of component fields and their data types were specified.

Historically, the first databases were located on magnetic tapes, which led to certain peculiarities of structure of the conceptual schema. In particular, the schema was hierarchical or networked.

A hierarchical schema is a tree-like structure, at the top of which the main record is located. This record is linked by references to subordinate records. These records, in turn, also contain pointers that link them to their subordinate records. Thus, the records make a hierarchy consisting

of several layers, and a programmer (or an application program) can follow this hierarchy, to move from one record to another, performing the so-called **navigation**.

Network schema is a next stage in development of schemas. At its highest level not one but several types of records are located, navigation can be performed from several different points and can be carried out both in descending and ascending directions of the hierarchy.

Both models are convenient for processing hierarchically arranged information, however certain difficulties arise when operating on data connected by complex logical connections.

Until the resources of computers were not so powerful and the number and demands of users were not so great, both approaches were more or less acceptable. But the development of technology has increased the speed and capacity of peripheral devices, multi-tasking and multi-user operating environments have emerged, and finally the number of users has increased. The issue of developing database schema optimization criteria was on the agenda.

At this stage, a special role belongs to Edgar Codd, who in 1970 published several papers on the conceptual modeling of data in large data banks, which qualitatively changed the current situation in this field and led to the creation of relational databases.

E. Codd was based on the idea that any conclusions about the navigation results should be made on the semantic analysis of the information placed (or to be placed) in the database.

Therefore, he transformed the conceptual schema of the database, establishing the notion of its normalization. By introducing some information redundancy in the records, it was possible to eliminate the hierarchy of subordination.

In a formalism developed by E. Codd, a relational database schema is defined as a finite set $R = \{A_1, A_2, ...A_n\}$ of attribute names. Each attribute A_i relates to its domain D_i - a finite or at most countable nonempty set of values.

A record (which is called a tuple in the theory of relational databases) is represented as some mapping t from the schema R to the set union of the domains $\bigcup_{i \in 1..n} D_i$. This mapping meets the requirement: $t(A_i) \in D_i$.

A set of tuples with an identical schema is called a relation.

If $T = \{t_1, t_2, \ldots, t_p\}$ is a relation and P denotes any subset of its schema, then $T(P)$ will be used to denote the set of reductions of the components of relation T on P.

Thus

$$T(P) = \{t_i(P) | i = 1..n\}$$

A subset of a schema that uniquely identifies a tuple within a relation is called the key to that relation. So $K \subseteq R$ represents a key of a relation if following condition is satisfied:

$$\forall t_1 \in S, \forall t_2 \in S: ((t_1(K) = t_2(K)) \Rightarrow t_1 = t_2)$$

Navigation within the database is considered as mapping from corresponding relation M to the Boolean 2^Q of target relation Q.

Thus, if M is a relation with the schema R_1 and N is a relation with the schema R_2 then the result of navigation from the tuple t of relation M to the relation N is represented as a function

$$F(t \in M) = \{l \mid l \in N \;\&\&\; ((R_1 \cap R_2 \neq \emptyset \wedge t(R_1 \cap R_2) = l(R_1 \cap R_2))\}$$

where $\emptyset$ denotes an empty set and && denotes the operation of logical multiplication. And since both the initial and final information are important in the navigation process, this mapping is considered as a binary operation on two data sets (called a Union operation in the relational

formalism and denoted mostly by a symbol $\otimes$) with the result as a set containing the information requested by a programmer (or an application program).

More precisely, if M is a relation with the schema R_1 and N is a relation with the schema R_2 then their union $S = M \otimes N$ is a relation with the schema $R_1 \cup R_2$, and this relation satisfies following condition:

$$S(R_1) = M \;\&\&\; S(R_2) = N,$$

which means that if the tuples of relations involved in the operation do not match with each other on the intersection of corresponding schemas, the result is an empty set. And if the intersection of those schemas is itself an empty set, the result of operation is undetermined, that in fact means that navigation can lead to any tuple of target relation.

Ultimately, the database was presented as an algebraic system, the objects of which are so called relations - the sets of formatted records or tuples. Operations on them led to data manipulation and (hence) navigation between relations. Two alternative representations of this algebraic system are known – the so-called Relational algebra and Relational calculus. Already E. Codd, in his time, proved their equivalence [2].

Based on mentioned mathematical formalism, software mechanisms were created, which provided programmers with tools for describing data structures in databases, to perform data allocation and data manipulation. One (but not the only) of the example of such a mechanism was the declarative language of data manipulation SQL. Due to its popularity, the mechanism of organization and access to information in databases, which is based on relational formalism, is mostly known as the SQL data access mechanism.

Finally, the notion of data model as a formal theory of data representation and data processing has been established in the database theory. This theory combines at list three different aspects – that of data structure, data manipulation, and data integrity. As for the organization of database, it combines three views – outer view, conceptual view and physical view.

Outer view represents the database configuration from the point of view of a user.

Conceptual view conveys the overall logical structure of the database represents a generalized model of a problem area.

Physical representation is under responsibility of database management system and operation environment.

Such an architecture ensures the isolation of data processing application from the logical and physical representation of the database.

This concept was first presented in 1978 by the American National Standards Institute (ANSI) and the Standards and Specifications Planning Committee (SPARC), and eventually became the foundation for understanding the basic functional characteristics of databases and database management systems.

As for the database management system (DBMS), it was eventually formed as a combination of software and language tools that provide:

1. Creation of a database and its joint processing by various users.
2. Protection of database information, its safe storage and adequate representation of a problem area (blocking, backup and restore).
3. High-level programming systems (languages and appropriate compilers) for development of user interface and information systems, dedicated to information processing and representation of results in textual, graphical and other formats.

The further development of computers and information technologies gave rise to two seemingly opposite trends – on the one hand, the integration of databases, and on the other hand, the possibilities of their distributed and parallel processing.

The integration was caused by the realization of the truth that the database represents an information model of a subject area or some of its fragment.

In a narrow sense, it can be understood as a problem arising from the division of information into operational and analytical data and therefore might be reduced to equipment of database management systems with adequate tools.

This refers to division of databases into OLTP and OLAP databases, including the problems of building data warehouses and data mining [3]. Here we already witness a deviation from the relational standard – examples of implementation of so called NOSQL mechanism.

In a broad sense, this problem implies the development of such methods of subject area modeling that will allow to manage more effectively the architecture and the lifecycle of its information resource.

In the 1970s, in the Muskhelishvili Institute of Computational Mathematics, within the framework of the project of the information system supporting industry statistics, a large data bank management system was built which ensured within the industry the systematic accumulation and management of analytical information resources.

As initial information the system received the regulatory reports of the functional unites of the industry and integrated them into macroeconomic indicators of the industry.

For data storage the system used a multidimensional information model, access to data was carried out through a hierarchy of indexes and the user had opportunity to apply the built-in formula analyzer for calculations to be performed in interpretation mode. This ensured independence of applied programs from the data structure. To optimize the modes of operation the system had a planning and management subsystem and, along with it, a semantic metadata management subsystem, which allowed the accumulation and analysis of the terminology in the subject area.

Calculations were possible both in batch mode and at the level of individual formula processing.

In the 1980s the mentioned system was implemented in the health sector of the former USSR as an information system for management of industry statistics [4].

In 1981-1985 the researchers from the Muskhelishvili Institute of Computational Mathematics were actively involved in development of network-type database management system “COMPAS” held in Computational Center of Academy of Science of USSR, particularly in design and implementation of metadata model of this system. The project was interested from the point of view that the meta-model represented the network scheme of the dictionary-reference database described and developed in terms of the COMPAS system itself on the basis of standard data description (DDL) and data manipulation (DML) languages developed by CODSYL for network-type databases [5].

In the last half century, the development of computer technology has replaced the earlier centralized single-processor systems with high-speed computer networks, or distributed systems. Today the modern global network allows millions of machines around the world to exchange the information at the speed from 64 kb to megabits per second. Management systems of parallel and distributed databases are nowadays the main tools for intensive data processing.

Attempts of reaching sustainable optimal position for subject areas are often faced with problems caused by presence of corrupted information resource, when information necessary for decision-making is not properly motivated and argued, sources of data - unreliable, information

obtained by the user - incomplete, inconsistent, irrelevant, etc. As a result, lack of clarity in cause-effect relationships, unawareness on the results of corrective actions is observed.

Making one of the major factors in social relations, analytical information becomes a social asset and turns into strategic resource of a society. Evidently, the problem of accumulation and utilization of this resource in industrial mode is actualized.

Industrial mode of information resource development implies possibility of its collection and completion on regular basis, its distribution among all stakeholders, with access to information limited only by the measure of its privacy and regulated (presumably) by economic and civil relations.

At the present stage of development of information technologies, for technical point of view, we may talk about building of a cyber infrastructure to create and maintain an information space for accumulation and distribution of shared analytical information resource of various subject areas.

Existence of such a repository could be regarded as one of the conditions for sustainable development of any country, as its presence leads to:

- Improvement of interaction within different subject area;
- accumulation and sharing of knowledge in various spheres of activity;
- transparency of administration and control;
- realization of opportunities for decision making on the basis of modern management techniques;
- prevention (or even elimination) of conflicts.

Corporate analytical information resource of a society is regarded as union of data clusters representing different subject areas (e.g. healthcare, education, industry, agriculture, etc). Solution involves, for each cluster, data acquisition, on some regular basis, from primary information providers of subject areas, its further processing and transformation into information aggregates, relevant to corresponding levels of subject area's organizational hierarchy, subsequent provision of information needs of different stakeholders, maintenance of long-term information archives.

Data collection according to coordinated models would ensure integrity and consistency of information resource, comparability and compatibility of data, transparency of information space.

Among the requirements to the cyberspace, as a guarantee of quality of information resource provided, following may be mentioned:

- Openness, meaning ability to include new sources of information and new points and layers of data aggregation;
- Compatibility and comparability of data, arriving from different points of environment to data processing nodes and end user;
- Transparency as the possibility to trace sources and origins of any data aggregates;
- Scalability, as ability to customize to different physical configurations – from standalone desktop computers to complicated networks.
- Security

Architectural decision should be flexible enough to be adaptable to the features of different communities, allowing them to organize their information space in suitable manner and to modify it easily in case of necessity.

These problems should be solved within the framework of the system approach.

References

[1] Ramakrishnan R., Gehrke J. Database Management Systems. Mc-Graw Hill Higher Education. 2003.
[2] Mayer D. Theory of Relational Databases. Computer Science Press. 1991.
[3] Kimball R., Ross M. The Data Warehouse Toolkit. John Willey and Sons, 2002.
[4] Glonti I. G. (1977). Ob odnom opite razrabotki informacionnoi bazi otrasli na primere zdravooxranenia. Akademia nauk GSSR, Vichislitelnyi centr, Trudi XVII:2. (in Russian).
[5] Bobrov A.V., Papiashvili M. R., Tikhomirov S. E. et all. (1983) Setevaya SUBD Kompas-BESM. Tezisi dokladov 2-i Vsesoyuznoi confrencii Banki Dannih, seqcia. 2, Kiev. str. 132-135ю (in Russinan)
[6] Tamer Ozsu M., Valduriez P. Principles of Distributed and Parallel Database Systems. Springer 2011.

Article received: 2022-08-09