UDC 621.37

CONCEPTS OF MODELING, PROCESSING AND DATA WAREHOUSING

N.A. Atroshchenko

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus,
natasha@atroshchenko.ru

Abstract: In the era of digitalization of various areas of society development, data has acquired the status of a vital resource. The constantly increasing volume of data in the world since 2000, both structured and unstructured, coming from different sources and having different formats, has caused the emergence of new tasks, required innovative non-standard solutions for fast processing, secure storage of data and well-organized transparent access to them. Today, various concepts exist in parallel and are successfully applied, often complement each other, and are successfully applied to solve emerging problems. A pressing issue is also the search for directions that allow satisfying the need for modeling and storing information.

Keywords: data structures, Big Data, data warehouses, DaaS, DWH, ORM systems.

I. INTRODUCTION

Our time is characterized by the growth of data in geometric progression. Today, information is collected in huge volumes from various sources: the Internet, contact centers, mobile devices, web cameras, IoT sensors, social media, etc. The data format may not have a clear definition: text documents, images, video, audio, data from social networks, server logs, financial transactions.

II. MAIN TEXT

Without a clear structure and orderliness, it is impossible to apply generally accepted concepts of automation and analysis to data.

In addition, the mere presence of data, even well-structured, will not be enough for their technological, effective maintenance. They must be transformed into an information structure ready for management, storage and modification, which is impossible without displaying them in a certain form or model. In this regard, it is important to decide on the choice of concepts for storage, modeling and processing.

A data warehouse (DWH) as an information system is nothing more than an infrastructure with the necessary tools for prompt access to them and processing. We are talking about the essence of data warehouses as structured, integrated, subject-oriented data sets with a time reference for their rapid processing and information support. The need to work with data consolidation, integration of multiple data sources from different systems (from files, external data sources, RDBMS, Data Lakes, etc.), the search for convenient, fast and secure ways to work with data is growing. This is the reason why issues related to the design of data warehouse architecture are so relevant and in demand today.

In this regard, an important aspect is the principle of organizing an object storage as a software system designed to manage large volumes of data used for forecasting, reporting, analytics, and making business decisions.

The next evolutionary stage in the development of the data storage concept can be called Enterprise Data Warehouse (EDW), or corporate data warehouse, providing a wider range for typification of values and sources of source information. Data presentation is characterized by a consistent, structured, standardized view, the absence of discrepancies for users.

Due to the explosive growth of the volumes of processed information, the terms "Big Data" and "Hadoop" appeared for areas of distributed processing of big data. With the advent of modern technologies, such as massive parallel processing (MPP) architecture for commercial DBMS and the development of cloud technologies, the concepts of Data Fabric and Data Lake have come forward. New technological directions based on process automation and the prospects of cloud storage have led to the active use of cloud services Data as a Service (DaaS) and architectural directions Data Mesh and Data Lakehouse.

The choice of the prevailing concept of the DWH management system, which includes relational, non-relational (NOSQL), object-oriented, XML-oriented, hybrid object-relational, distributed - is an important aspect of building the architectural foundations of IT projects. Of course, it is important not to make a mistake in choosing the presentation of the data structure.

Until a certain time, the relational approach to data modeling was considered more traditional; it is traditionally more often represented in the corporate world and uses the widely used standard SQL language. Relational modeling initially involves the development of ER diagrams that define the structure of entities and relationships, as well as visualization of the relationships between them [1].

The definition of relationships between tables and tabular data is carried out by setting foreign and primary keys. The presence of a primary key guarantees unique identification of table records. A foreign key is needed to establish relationships between tables and is a column of a child table that refers to the primary key of the parent table. The required form of databases is achieved through 6 forms of normalization. This way, data is structured, retains integrity, access to it is simplified, and the redundancy of stored information is reduced.

Using object-oriented concepts in building data models (encapsulation, inheritance, polymorphism) significantly simplifies development and design. Object-relational DBMS combine the capabilities of object-oriented and relational databases. This provides a number of advantages. For example, one of the options for solving the problems of structuring data coming from various sources organized using different metamodels (Metadata Impedance Mismatching, is the creation of an object-relational mapping. In practice, the possibility of including an intermediate layer in information systems that has a ready-made API with a full set of working interfaces for accessing data has proven itself to be quite good in this regard.

Of course, ORM (Object-Relational Mapping) systems, which are based on the mechanism of linking a database management system with an object-oriented programming language, have all the priorities of object-oriented access to data. Such systems have long been successfully used for efficient processing, access and storage of data. Many mechanisms have been developed that eliminate the problems of data typing and their transformation for storage in a DBMS.

In this regard, it is possible to emphasize the advantages of object-oriented DBMS, characterized by the provision of natural object-oriented interfaces for projects in high-level implementation languages.

A sufficient number of ORM systems have been developed to date, successfully performing tasks of efficient and productive work with data storage. A kind of virtual DBMS is created. In this case, the developer must provide (generate) mapping of DBMS tables to solution classes, and entities (Entity Framework) will perform the necessary transformations.

Among the indisputable advantages of ORM systems are: reduced development time, transparency of maintenance, no need to write most SQL queries, a standardized and understandable maintenance process, resistance to internal and external errors. A classic software project in a high-level language, as a rule, also includes layers of storage, processing, analysis and visualization of data.

Distributed databases [3] can be called the next step in the evolution of data storage and processing. The impetus for their emergence was the branching and widespread use of computer networks. This is the most promising direction of development of data processing and storage. However, it requires solving many problems.

Along with issues of support, configuration, organization of request processing, fault tolerance and scaling of applications, there are issues of reliability, performance, transaction synchronization, problems of communication and computing resources, management of data storage as a single entity, etc.

III. CONCLUSION

The search for solutions that meet modern requirements for data processing and storage is currently being conducted in all directions. Ways to improve the performance of data access, harmonization of data structures coming from various sources, organization of storage security, elimination of conflicts, collisions, variability of typing, etc. are successfully solved using various approaches, innovative and complex developments. Their implementation in modern information systems opens up prospects for solving problems of data storage and processing at the global level.

REFERENCES

[1] Information Modeling and Relational Databases / Terry Halpin: Morgan Kaufmann, 2024. 1084 Pages.

[2] The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling / Ralph Kimball, Margy Ross: Wiley, 2013, 3rd Edition. 608 Pages.

[3] Cloud Data Warehousing Volume I: Architecting Data Warehouse, Lakehouse, Mesh, and Fabric/ Barry Devlin: Kindle Edition, 2023. 189 Pages.

КОНЦЕПЦИИ МОДЕЛИРОВАНИЯ, ОБРАБОТКИ И ХРАНИЛИЩ ДАННЫХ

Атрощенко Н.А.

Белорусский государственный университет информатики и радиоэлектроники,
Минск, Республика Беларусь, natasha@atroshchenko.ru

Аннотация: В эпоху цифровизации различных направлений развития общества данные приобрели статус жизненно важного ресурса. Постоянно возрастающий в мире с 2000 г. объём данных, как структурированных, так и неструктурированных, поступающих из разных источников и имеющих разный формат, вызвал появление новых задач, потребовал инновационных нестандартных решений для быстрой обработки, безопасного хранения данных и хорошо организованного прозрачного доступа к ним. Сегодня параллельно существуют и успешно применяются различные концепции, часто взаимодополняют друг друга, успешно применяются для решения возникающих задач. Злободневным вопросом также является поиск направлений, позволяющих удовлетворить потребность в моделировании и хранении информации.

Ключевые слова: структуры данных, Big Data, хранилища данных, DaaS DWH, ORM-системы.