TECHNICAL UNIVERSITY OF GABROVO

Faculty of Electrical Engineering and Electronics

M. Eng. Velin Sabinov Hadzhiev

Modelling of Data Structuring, Storage, and Processing Operations on the Internet

A B S T R A C T

of the dissertation

for the award of the educational and scientific degree of PhD

Field of higher education: 5. Technical sciences

Professional field: 5.3. Communication and Computer Engineering

PhD Program: Automated Information Processing and Control Systems

Scientific supervisor: Assoc. Prof. Dr. Eng. Aldeniz Enverov Rashidov

Reviewers: 1. Prof. DSc Eng. Raycho Todorov Ilarionov

2. Assoc. Prof. Dr. Eng. Galina Ivanova Ivanova

Gabrovo, 2024

The doctoral dissertation has been discussed and recommended for official presentation at a meeting of the Extended Departmental Council of the Department of "Automation, Information and Control Systems" at the Faculty of Electrical Engineering and Electronics of the Technical University of Gabrovo, held on December 11, 2024.

The doctoral dissertation consists of 156 pages. The scientific content is presented in an introduction, four chapters, and a conclusion, and includes 32 figures and 13 tables. A total of 118 literature sources are cited. The numbering of figures, tables, and formulas in the abstract corresponds to that in the dissertation.

The research for the doctoral dissertation was conducted in the Department of Automation, Information, and Control Systems at the Faculty of Electrical Engineering and Electronics, Technical University of Gabrovo.

The official presentation of the doctoral dissertation will take place on March 12, 2025, at 1:00 PM in room 3410 at the Technical University of Gabrovo.

The materials related to the presentation are available for interested parties in office 3209, Building No. 3 (Rectorate) of the Technical University of Gabrovo.

The reviews and opinions of the members of the academic jury, as well as the abstract, are published on the university's website: www.tugab.bg.

Author: © Velin Sabinov Hadzhiev
Email: hadjo75@abv.bg
Title: Modelling of Data Structuring, Storage,
and Processing Operations on the Internet
Print Run: 8 copies (in English)
Place of Printing: University Publishing House "Vasil Aprilov"
at the Technical University of Gabrovo

GENERAL CHARACTERISTICS OF THE DOCTORAL DISSERTATION

Relevance of the Problem

In their daily activities - scientific, research, business processes, and more - people continuously generate data that often needs to be stored in a specific format and schema to later be used for processing and various purposes. Storing data in a specific schema implies arranging and storing them in a manner that facilitates its processing to extract dependencies and valuable information. Organizing and storing data within a specific schema involves operations related to structuring, storing, and processing this data, which are always performed by specialists.

The dissertation addresses the current scientific problem of finding a solution that would allow any Internet user generating or processing data to describe the data storage structure at will, as well as to perform manipulations on this data (input, editing, sharing, etc.).

Aims and Objectives of the Dissertation

The aim of this dissertation is the modelling of data structuring, storage, and processing operations on the Internet.

In accordance with the stated aim, the following main objectives are addressed:

1. Modify and investigate methods for structuring, storing, and processing data.
2. Develop and investigate algorithms for structuring, storing, and processing data.
3. Develop models of operations for structuring, storing, and processing data on the Internet.
4. Investigate and compare the proposed models.
5. Identify the factors that characterize information systems as systems for structuring, storing, and processing data on the Internet.
6. Apply the developed models in the design and development of systems for structuring, storing, and processing data on the Internet.

Research Methods

The research methods used to address the tasks set in the dissertation are: theoretical analysis, computer-aided design, modelling, and simulation studies, with selected methods for structuring, storing, and processing data applied for this purpose. Modelling of processes for structuring, storing, and processing data on the Internet has been carried out, including simulation of the operation of a hybrid model for managing distributed data and its interaction with the global network. An analysis of the effectiveness and cost-effectiveness of implementing such models has been performed, contributing to the improvement of the quality and accessibility of the information infrastructure for users.

Applicability

The applicability of the doctoral dissertation is related to the development of a data operation optimization method that integrates best practices and techniques for data structuring, storage, and processing. The effectiveness of the proposed hybrid model has been

demonstrated through simulations and tests in real-world conditions, evaluating its performance, resilience, and scalability. Based on this method, a web-based data management system has been developed, providing access to databases for a wide range of users. Practical scenario tests confirm its high efficiency and applicability in real-world conditions.

Approbation of the Doctoral Dissertation

The main results presented in the doctoral dissertation have been published in nine papers in international conferences and scientific journals, with four of them abroad and five in Bulgaria. One of these publications is a sole-author paper. The number of these publications fully meets the minimum requirements regarding this criterion. Two of the works were presented at the International Scientific Conference "Automatics and Informatics, 2019", and both were published in the journal "Journal Automatica and Informatics, 2019", issues 2 and 3. One publication was presented at the international scientific conference ICAI, 2021, Varna, Bulgaria, three were presented at the international scientific conference ELECO, Bursa, Turkey, during the period 2019 – 2021, and one at the international scientific conference ICCCNT 2022, Mandi, India, which are indexed in Scopus with SJR. The publications presented a significant part of the conducted research and outlined the main conclusions of the doctoral dissertation.

The results achieved from the doctoral dissertation were reported at various scientific conferences. Many of the presentations were co-authored with the scientific supervisor. Parts of this work and the dissertation as a whole were repeatedly discussed in the Department of Automation, Information, and Control Systems at the Technical University of Gabrovo.

Structure and Volume of the Doctoral Dissertation

The doctoral dissertation includes: an introduction, table of contents, four chapters, classification of contributions, publications related to the dissertation, and literature. The total volume of the dissertation is 156 pages, which includes 32 figures and 13 tables. The list of used literary sources contains 118 titles.

CONTENT OF THE DOCTORAL DISSERTATION

CHAPTER I: OVERVIEW AND ANALYSIS OF EXISTING MODELS AND METHODS FOR DATA STRUCTURING, STORING, AND PROCESSING ON THE INTERNET

1.1. The topic relevance

In their daily activities - scientific, research, business processes, and more - people continuously generate data that often needs to be stored in a specific format and schema for later processing and various purposes. Collecting data in a specific schema implies arranging and storing them in a way suitable for data processing to extract relevant insights and valuable information. Arranging and storing data in a specific schema involves performing operations related to structuring, storing, and processing these data, all of which are typically carried out by specialists.

The present doctoral dissertation is dedicated to an important scientific problem - finding a solution that allows each Internet user, whether generating or processing data, to freely describe the data storage structure and to perform manipulations on these data (such as input, editing, sharing, etc.).

1.2. Overview of existing models, methods, and architectures for structuring, storing, and processing data on the Internet

Normalized Data Factory Modelling: This method was introduced by Bill Inmon based on the concept of data normalization [27].

Figure 1.1 presents an example of a normalization process to the Third Normal Form, which underpins the method proposed by Bill Inmon.

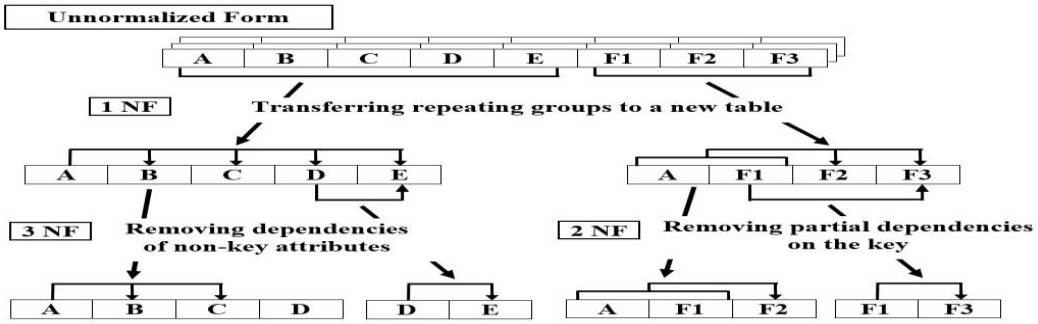


Figure 1.1 Example of a normalization process to the third normal form

The purpose of the normalization process is to eliminate redundant relationships, which increases efficiency and avoids anomalies in data storage.

Dimensional Modelling: This method was introduced by Ralph Kimball [27]. In dimensional modelling transaction and reference data are isolated in fact-tables and tables with dimensions (dimension tables). Figure 1.2 illustrates the schemas commonly used in building data storage systems.

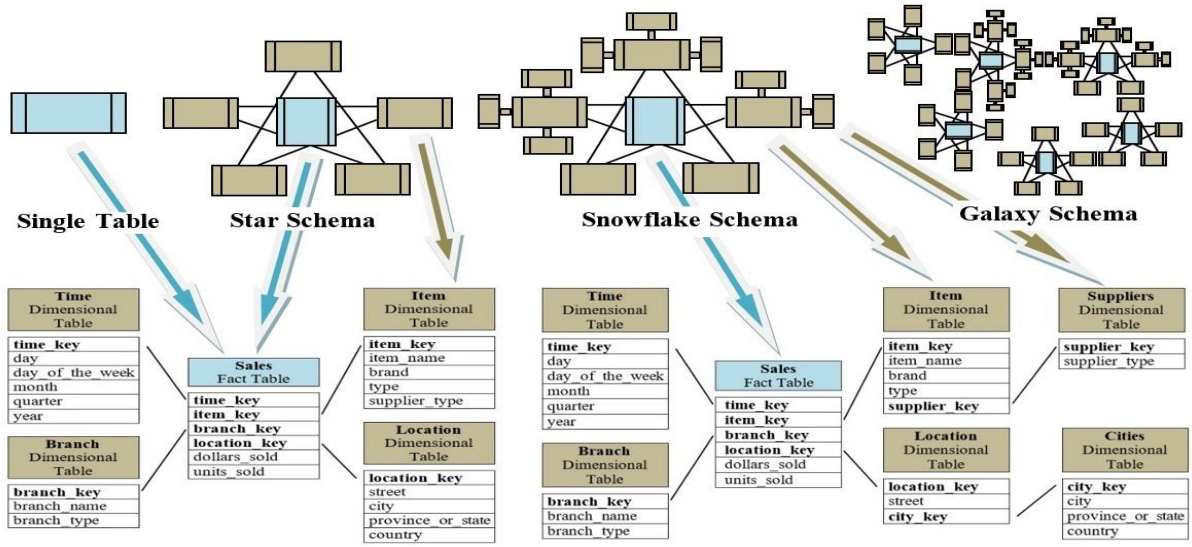


Figure 1.2 Basic dimensional modelling schemas

Table 1.1 summarizes the main differences between star schema and snowflake schema. These differences highlight the main advantages and disadvantages of each schema depending on the specific needs and requirements of a given project or organization.

Table 1.1 Differences between "Star Schema" and "Snowflake Schema"

Feature	Star Schema	Snowflake Schema
Data redundancy	There is data redundancy due to the denormalized dimensions.	Reduced data redundancy due to normalization.
Data management and storage	Easier to manage and maintain. Requires more space to store the dimensional tables.	Due to the complexity, it is more difficult to manage and maintain. Requires relatively less space to store the dimension tables.
Normalization	It contains denormalized dimension tables that are in the Second Normal Form (2NF).	It contains dimensional tables, normalized to the Third Normal Form (3NF).

Feature	Star Schema	Snowflake Schema
Structure	A central fact table encircled by denormalized dimension tables.	A central fact table encircled by hierarchically linked normalized dimensional tables.
Complexity	Simple structure, easy to understand and use. Queries represent direct joins between facts and dimensions to retrieve data.	Due to the normalization of dimensional tables, the structure is complex. Queries involve complex joins between facts and dimensions to retrieve the data.
Performance	Due to fewer relationships between tables, less time is required to execute queries.	Due to the additional relationships between tables, more time is required to execute queries.
Building and application approach	Top-down approach. Suitable for small to medium-sized data storage repositories.	Bottom-up approach. Suitable for large data storage repositories where normalization is needed.

Data Vault Modelling: It was introduced by Dan Linstedt [30][31]. The method is based on grouping the incoming data in the form of a hub, connection or satellite, using their tendency to change over time. Figure 1.3 illustrates the main elements involved in building data storage systems using the Data Vault methodology, along with their relationships.

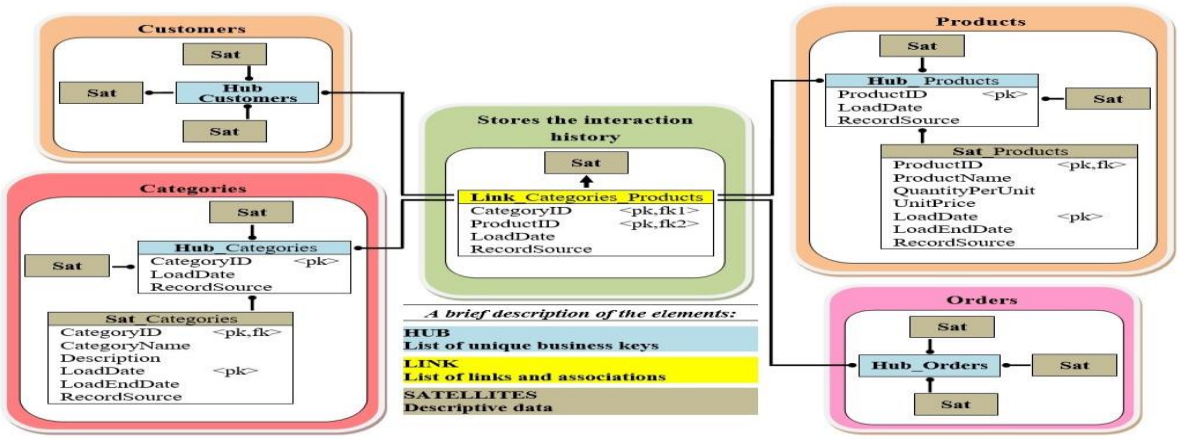


Figure 1.3 Basic data vault modelling elements

Hybrid Modelling: The method is a combination of dimensional architectures and architectures in 3NF, including an operational data store (ODS) with normalized data and dimensional data marts.

Data Lake Modelling: This method was introduced by James Dixon of Pentaho, who first used the term "Data Lake" in October 2010 [36][37]. The approach allows storage of large volumes of "raw" data in its native format, as well as semi-structured and structured data, with on-demand access to it [38]. Figure 1.4 shows the main components of data lake-based storage systems and the connections between them.

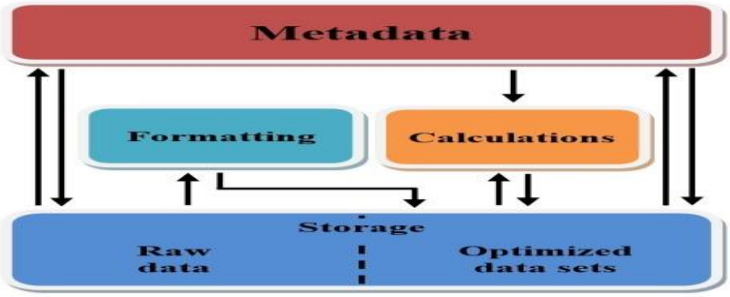


Fig. 1.4 Basic data lake modelling elements

The advantages and disadvantages of the five examined methods for data structuring and storage are summarized in Table 1.2.

Table 1.2 Comparison of methods for data structuring and storage.

Comparison Criteria	**Data structuring and storage methods**				
	Normalized Data Factory Modelling	Dimensional Modelling	Data Vault Modelling	Hybrid Modelling	Data Lake Modelling
Understandability, interpretation, and usage	- Good knowledge in the field of databases is needed.	+ Basic knowledge of databases is needed.	+ Basic knowledge of databases is needed.	It depends on the number and complexity of the methods used to provide a comprehensive solution.	- It needs professional skills, good knowledge, and readiness to work with large volumes of data.
Data duplication	+ Reducing duplication increases the consistency of the data.	- The lack of hierarchical separation of tables Increases data redundancy.	- Doubling the number of tables as opposed to 3NF modelling.	It depends on the number and complexity of the methods used to provide a comprehensive solution.	- A constant monitoring of data management and integrity maintenance is required. Otherwise, the data lake may turn into a data swamp.
Search and query performance	+ Quick index creation increases search speed. - Difficulties and delays in query execution arise due to the data being stored in more than one table.	+ Using of simple joins to retrieve the data increases performance. Quick index creation increases search speed. - Increasing the number of dimensional tables decreases the efficiency of query execution.	- In this modelling method, the performance of searching and data retrieval is not a priority.	It depends on the number and complexity of the methods used to provide a comprehensive solution.	- The performance of searching and data retrieval depends on the complexity of the algorithms used for data transformation before they are used for reporting or analysis.
Applicability in online data processing systems	+ Suitable for Online Transaction Processing (OLTP).	+ Suitable for Online Analytical Processing (OLAP).	+ The rules for loading hubs, links, and satellites have low complexity.	It depends on the number and complexity of the methods used to provide a comprehensive solution.	+ High efficiency when using tools for analyzing large volumes of data.
Flexibility	+ It allows logical grouping of similar or related data following the same schema.	- It is not suitable for long-term implementation if frequent changes in business requirements are needed.	+ Separate management of business keys from all business entity attributes, grouping them into hubs, links, and	It depends on the number and complexity of the methods used to provide a comprehensive solution.	+ Due to the possibility of horizontal scaling, large volumes of data can be stored and processed. Open-source technologies are

Comparison Criteria	Data structuring and storage methods				
	Normalized Data Factory Modelling	Dimensional Modelling	Data Vault Modelling	Hybrid Modelling	Data Lake Modelling
			related satellites. All attributes are managed as slowly changing dimensions, similar to dimensional modelling.		used, which are more flexible and cost-effective than proprietary technologies.
Complexity	- The complexity of the database increases, as the number of tables increases.	- It is necessary to use complex joins in query execution as the number of tables to be added increases.	+ Each relationship between business entities within the system is modeled as many-to-many.	It depends on the number and complexity of the methods used to provide a comprehensive solution.	- It involves large volumes of data, which complicates maintenance, management, and usage.

Conclusion: The main objective of the dissertation is to propose a solution that enables users to structure, store, and process data according to their requirements. The data should be stored in a tabular format with a defined structure and represented with their atomic values.

Designing and building an efficient data warehouse requires the application of an appropriate data storage model. The two main widely recognized models are the Inmon model and the Kimball model.

Bill Inmon model: Figure 1.5 shows a cloud database model, in which the Bill Inmon model was used to construct the data warehouse. Loading of the data warehouse is done by online transaction processing systems (OLTP). Data warehouse is in 3NF. Data marts are provided outside the data warehouse and new ones can be created when needed. They are in 3NF and involved in building data cubes that are used for online analytical processing (OLAP).

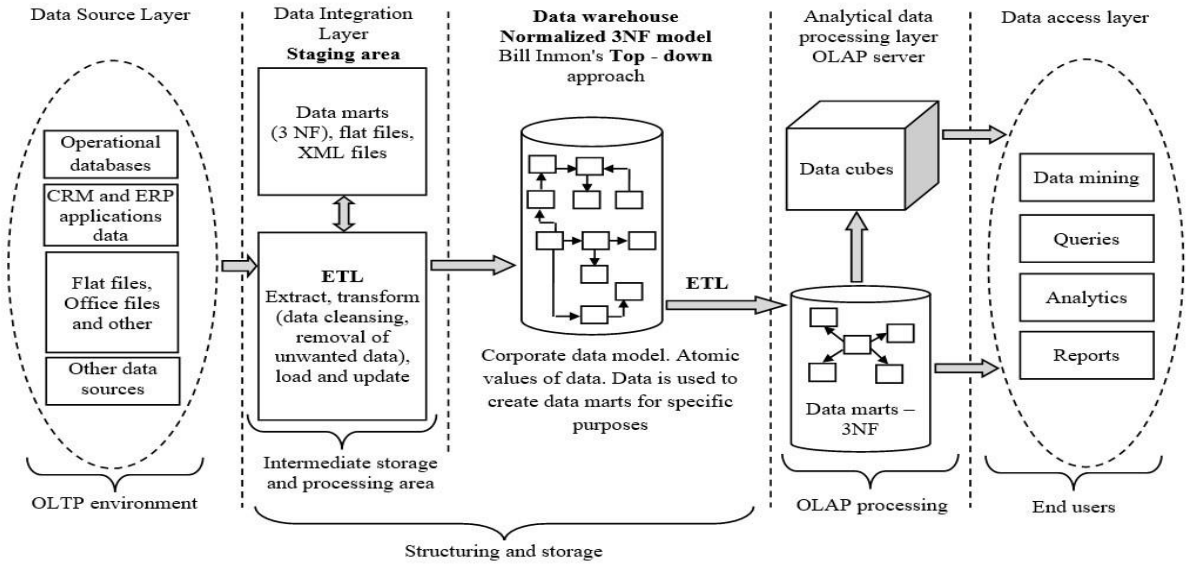


Fig. 1.5 Cloud database model based on Bill Inmon's model

Ralph Kimball model: Figure 1.6 shows a cloud database model, in which Ralph Kimball's model was used in the construction of the data warehouse. The data warehouse is built by data marts. Data marts are loaded with data from OLTP systems, which are typically relational databases in 3NF.

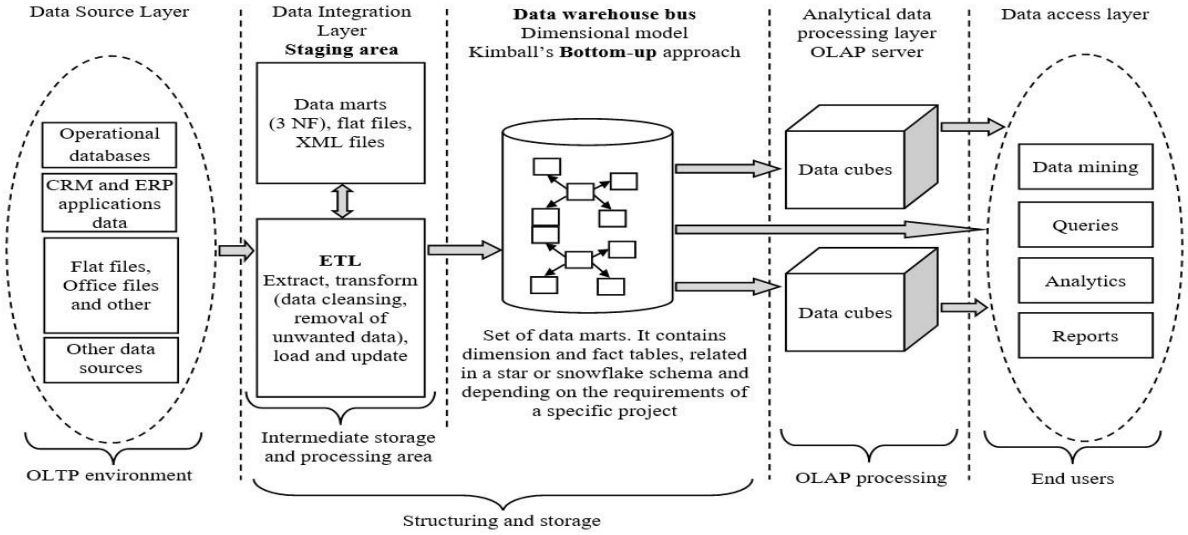


Fig. 1.6 A cloud database model based on Ralph Kimball's model

Data Vault Model: Figure 1.7 presents a model of a cloud database, based on the Data Vault Model introduced by Dan Linstedt. The three layers of the model ensure the isolation of the raw data storage from end users and define the various data extraction layers.

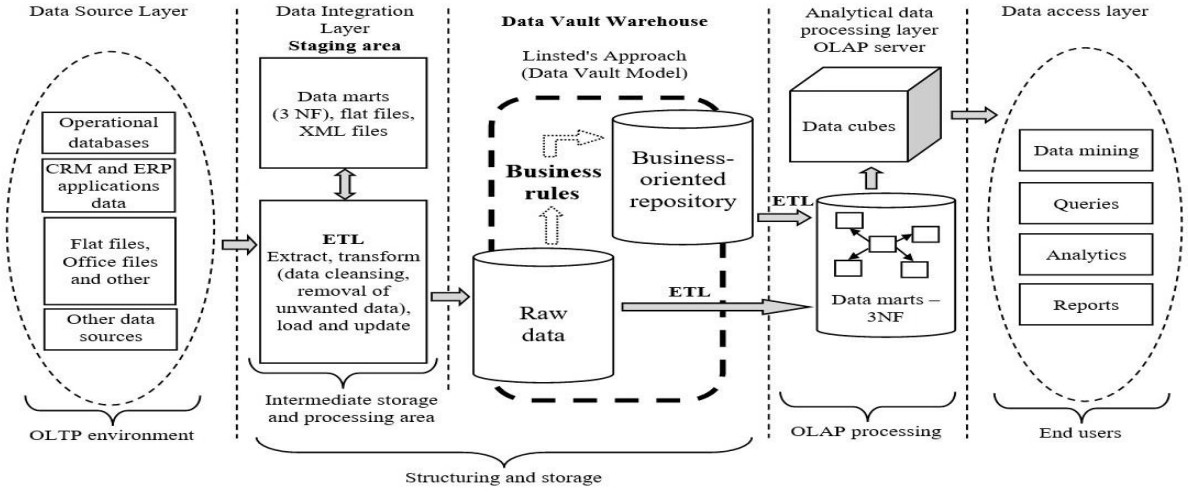


Fig. 1.7 A cloud database model based on Dan Linstedt's model

Table 1.3 Basic Differences between the Models of Inmon, Kimball, and Linstedt

Feature	Inmon	Kimball	Linstedt
Ideology.			
Target audience.	IT Specialists.	End Users.	End Users.
Role in the organization.	An integral part of the Corporate Information System (CIS) for Business Intelligence and Business Management.	Transforms and stores operational data.	Stores raw data and other data after business rules have been applied to them.

Feature	Inmon	Kimball	Linstedt
Objective.	Providing a stable technical solution based on proven methods and technologies for data structuring, storage, and processing.	Providing a solution that allows end users to initiate queries to the database, with response times within acceptable limits.	Providing a secure and comprehensive solution based on proven methods.
Methodology and architecture.			
Development approach.	Top-down	Bottom-up	Hybrid Approach: A combination of 3NF and Star Schema.
Purpose.	Corporate data warehouse storing the atomic data values. Provides data for the databases of individual departments.	Data marts modelling business processes separately. Enterprise consistency is achieved through a data bus and appropriate dimensions.	The Data Vault contains the business-oriented warehouse and the raw data warehouse, which provide data to the data marts.
Complexity.	Quite complex.	Fairly simplified.	Fairly simplified.
Development and evolution methodology.	Spiral Development and Evolution Methodology.	A four-step process for dimensional modelling, distinguishing it from relational database methods.	Based on an agile methodology focused on rapid development and implementation.
Implementation costs.	High costs during the development and implementation phase. Lower costs for expansion and maintenance.	Low costs during the development and implementation phase. Higher costs for expansion and maintenance.	Lower costs compared to the other two approaches.
Implementation duration.	Long implementation duration.	Short implementation duration.	Short implementation duration.
Physical design discussion requirements.	Fairly comprehensive.	Rather superficial.	Rather superficial.
Data integration.			
Integration of multiple sources.	The transformation rules must be applied during data extraction and loading.	The transformation rules must be applied during data extraction and loading.	Separating contextual data from business keys reduces the complexity of data extraction and loading.
Complexity of data extraction, transformation, and loading processes.	Low complexity of transformation rules if the data model closely resembles the models of the data sources.	The transformations between the online transaction processing model and the dimensional model are complex.	Low complexity of the rules for loading hubs, links, and satellites.
Data Modelling.			
Data orientation.	Theme-based, data-driven.	Oriented towards a specific process.	Data-driven and process-oriented, focused on a specific process.
Implementation tools.	Traditional: entity-relationship diagrams and data sets (ERDs, DISs).	Dimensional modelling, distinct from relational modelling.	Hubs, links, and satellites modelling.

Feature	Inmon	Kimball	Linstedt
End user accessibility.	Low end-user accessibility.	High end-user accessibility.	High end-user accessibility.
Life cycle management.			
Flexibility in changing the data source model.	Changes in the tables needed.	Frequent changes in the data source model affect the data warehouse model.	The existing tables will not be affected. The only change is the addition of the corresponding satellites.
Flexibility in changing analysis requirements.	The model needs to be modified if the required data is not available in the data warehouse.	The changes in requirements affect the data model.	No changes to the model are required. Only the data delivery to the data marts needs to be adapted.
Easy model change.	In some cases, historical data needs to be migrated.	In some cases, it is necessary to make changes to certain tables.	The only change is the addition of the corresponding satellites.
Audit and traceability.	Chronological collection of information by inserting a new record for each change.	It uses the concept of "slowly changing dimensions" to collect information chronologically.	Chronological collection of information by inserting new links and satellites.
Query execution efficiency.	Query execution is very slow due to the normalized 3NF data structure.	The model is designed to be highly efficient for intensive queries due to the denormalized nature of the dimension tables.	Due to the high standardization of the data, direct queries are very slow. Dimensional data marts are required for analysis and reporting.

The choice of a model for structuring and managing data should be based on the specific needs of the organization, data requirements, and resources. Cloud data warehouses are a collection of databases and mechanisms for accessing data through a unified object [55]. To build a data warehouse, several computers can be used, among which the data is distributed.

Figure 1.8 shows the classic architecture of a data warehouse [57][58]. In this architecture, end users have direct access to the data in the data warehouse, which is sourced from multiple origins.

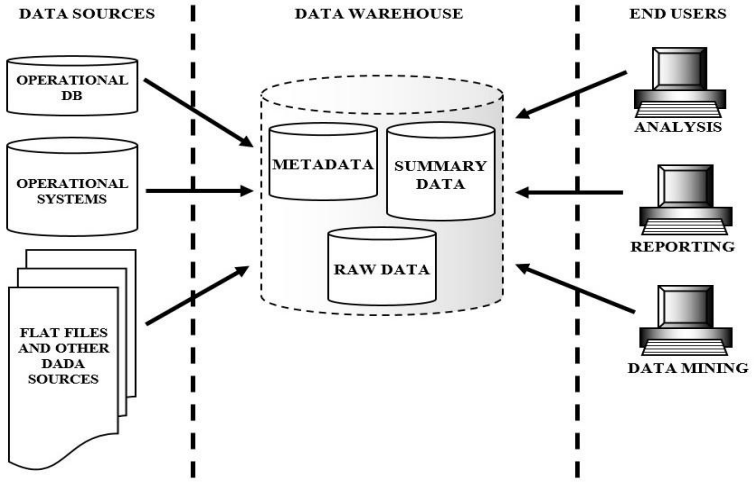


Fig. 1.8 Classical data warehouse architecture

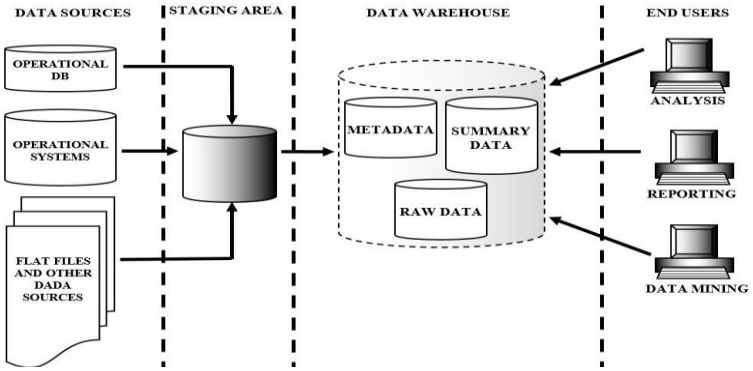


Fig. 1.9 Data warehouse architecture with a Staging Area

Although the architecture in Figure 1.9 optimizes the performance of the data warehouse during data integration, this may prove insufficient if a large number of users are working with the data simultaneously.

This disadvantage is addressed by expanding the architecture [57] presented in Figure 1.9, by adding data marts (Figure 1.10).

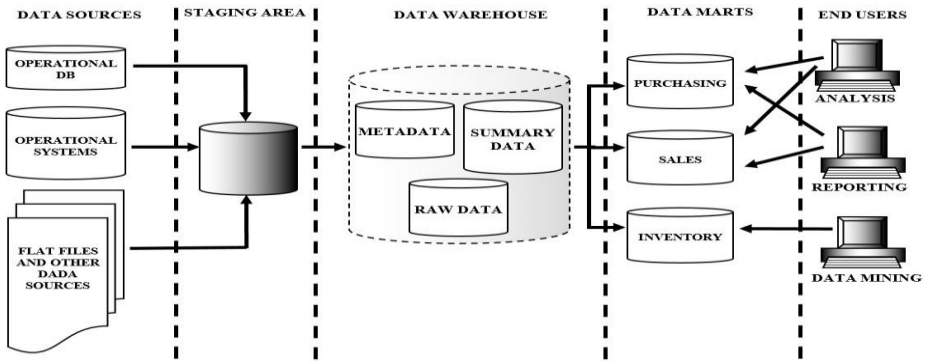


Fig. 1.10 Data warehouse architecture with a Staging Area and Data Marts

Figure 1.11 shows the architecture of a method for data structuring, storage, and processing in the cloud by any user on the Web [18]. The architecture supports an unlimited number of relational databases located on servers (Si), which are situated at key points – regions (Ri) in the global network.

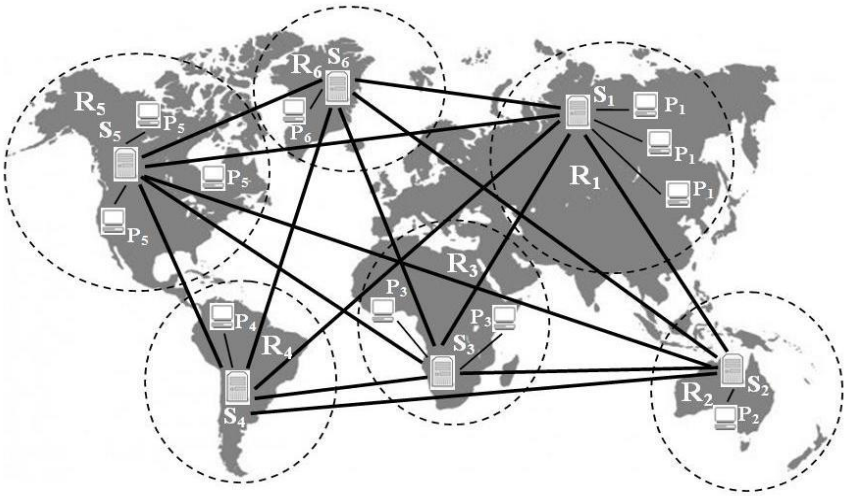
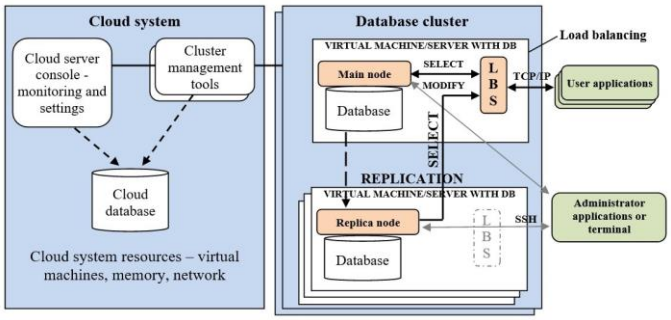


Fig. 1.11 Architecture of a method for data structuring, storage, and processing in the Web



The databases presented in Figure 1.12 are the fundamental unit in building the data warehouse, which is a relational database designed more for query execution and analysis than for transaction processing.

Fig. 1.12 Architecture of a cloud database

The data structuring, storage, and processing models can be created using methods such as: "Normalized Data Factory Modelling", "Dimensional Modelling", "Data Vault Modelling", and so on. Models can also be obtained by combining some of these methods through "Hybrid Modelling" or by altering the parameters of the algorithms used to generate the models.

The evaluation of data structuring, storage, and processing models can be performed using a SWOT analysis. Its purpose is to summarize the strengths and weaknesses, opportunities, and threats associated with the use of each chosen model, while considering several aspects that must be taken into account when selecting the correct data model. These aspects vary depending on the stage of the data lifecycle for which a given model is designed.

Table 1.4 summarizes the importance of different factors at each data lifecycle stage.

Factor	Creation	Storage	Analysis	Archive
Speed and frequency of data creation and modification.	High write speeds are required to ensure faster transaction execution. Data generated by end users or automated systems.	Moderately high write speeds are required. Large volumes of data that need to be stored sequentially.	Moderately high write speeds are required. Data aggregation may be required for efficient reporting.	Support for lower speeds. The archiving process is more reliable at lower speeds.
Data retrieval speed.	The data can be retrieved as soon as it is written. The granularity of the retrieved data can be the same as the inserted data, with minimal transformations.	Periodic data retrieval for generating smaller datasets. It may require data transformation and aggregation.	Data retrieval is necessary for reporting or management purposes. To satisfy the end user, data retrieval should be faster, and repetitive transformations and aggregations should be part of the data storage and model.	Data retrieval is only necessary in exceptional cases (audits, disaster recovery, etc.). Consistency with the existing datasets is more important than the speed of data retrieval.
Properties of Atomicity, Consistency, Isolation, and Durability (ACID).	The creation of data is part of transactions involving multiple steps. Compliance with ACID	Meeting all ACID requirements is not critical, but consistency of the dataset is expected before and after batch operations.	Meeting all ACID requirements is not critical, but consistency of the dataset is expected before and after batch operations.	Meeting all ACID requirements is not critical, but consistency of the dataset is expected before and after batch operations.

Factor	Creation	Storage	Analysis	Archive
	requirements is crucial for ensuring transaction consistency.			
Business scope.	Focused on a specific business activity.	Focused on multiple business functions or the entire enterprise.	Focused on specific requirements for reporting, artificial intelligence (AI), or machine learning (ML).	The scope depends on the stage of the data lifecycle (Creation, Storage, Analysis).
Access to the lowest level of data.	In most cases, if a system generates data with the lowest level of detail, access to the same level of detail is also required.	It may be necessary to store old data with different levels of detail. Access to the lowest level of data is important.	It may be necessary to store old data with different levels of detail. Access to the lowest level of data is important.	It may be necessary to store old data with different levels of detail. Access to the lowest level of data is important.

The most commonly used methods for evaluating data structuring, storage, and processing models include SWOT analysis. The goal of this analysis is to highlight the strengths and weaknesses, opportunities, and threats associated with the use of these models.

SWOT analysis of the Inmon model

Strengths	Weaknesses
• The data warehouse serves as the central repository, providing the foundation for building data marts that users interact with. This approach ensures data integrity is preserved; • Anomalies during data updates are minimized through reduced data redundancy. This simplifies ETL processes, making them more efficient and less prone to failure; • Business processes are easily understood, as the model provides a detailed representation of business entities; • Changes in business requirements or data sources are easily accommodated, as each element is stored uniquely in a single location within the data warehouse. This enhances the model's flexibility.	• As the number of tables increases, the model and its implementation can become more complex; • Experts in data modelling and business processes are required; • Building and initial setup take a significant amount of time; • Data marts are built and populated with data from the data warehouse; • It requires a relatively large team of specialists for management and maintenance.
Opportunities	**Threats**
• It is oriented towards corporate needs; • Three-tier architecture: data warehouse, data marts, data cubes, and data duplication; • It can be used to generate various reports to meet the needs of the enterprise.	• Increasing the number of tables makes data selection more difficult; • Difficulties in finding experts in data modelling and business processes; • ETL processes need to run longer, which will require more resources.

SWOT analysis of the Kimball model

Strengths	Weaknesses
• Enables rapid construction and deployment during the initial phase of the project; • The star schema is easy for business users to understand and straightforward to use for reporting. Additionally, it is well-supported by most business analytics tools; • The storage environment requires minimal space, making system management significantly easier; • The star schema model is highly effective because the database management system performs the "star join", which creates a Cartesian product by combining all dimension values and placing the fact table at the end of the selected rows. This is considered a very efficient database operation; • A small team of developers and designers is sufficient to maintain the efficiency of the data warehouse.	• The data is stored in data marts, which can lead to a lack of integration; • There is data redundancy; • Adding columns to the fact table can reduce performance, as the size of the fact table increases. This makes it more challenging to modify the dimensional model as business requirements evolve; • The model is focused on business processes rather than the entire enterprise; • High complexity in integrating legacy data into the data warehouse.
Opportunities	**Threats**
• It is focused on individual business processes, with users being actively involved; • Optimizing data extraction and management processes through a two-tier architecture: data marts and data cubes; • It is effective in tracking key performance metrics because data marts are focused on business process reporting; • High performance when using business analytics tools that require querying across multiple star schemas to generate reports. This can be effectively achieved using appropriate dimensions.	• It violates the principle of a "single source of information"; • Anomalies occur during data updates over time; • Difficulties in preparing certain enterprise reports.

Conclusion: Based on the SWOT analysis, it can be concluded that the Kimball model, with its bottom-up approach, is more scalable and allows for starting with small projects that can gradually expand. This approach ensures a quicker return on investment. In contrast, the Inmon model is more structured and easier to maintain, but it takes longer to build. A significant advantage of the Inmon model is that the Data Warehouse (DW) is designed in 3NF, which greatly facilitates the development and utilization of data extraction models.

SWOT analysis of classic data warehouse architecture

Strengths	Weaknesses
• Provides a unified data model, regardless of the diverse sources from which the data is obtained.	• The operational data in the warehouse are heterogeneous, with some of them potentially being redundant.
Opportunities	**Threats**
• Facilitates data processing and the development of the data warehouse.	• Due to the lack of mechanisms for removing redundant data, their volume may increase significantly, which could complicate operations and maintenance; • The occurrence of anomalies when updating data over time; • Difficulties in report generation.

SWOT analysis of the data warehouse architecture with a Staging Area.

Strengths	Weaknesses
• Provides a unified data model, regardless of the diverse sources; • Data integration through online transaction processing (OLTP) systems; • Data extraction from sources is carried out without delaying their operational systems.	• Difficult access when the warehouse is used simultaneously by a large number of users.
Opportunities	**Threats**
• Provides the ability for easier data processing and development of the data warehouse; • Enables the identification and resolution of data format inconsistencies.	• When a large number of users access the data warehouse simultaneously, the response time for user-generated queries may significantly increase or lead to system unavailability.

SWOT analysis of the data warehouse architecture with a Staging Area and Data Marts modules.

Strengths	Weaknesses
• Provides a unified data model, regardless of the heterogeneous sources from which the data is obtained; • Data integration through online transaction processing (OLTP) systems; • Data extraction from sources occurs without affecting the functioning of their operational systems; • Balanced load when the data warehouse is used simultaneously by many users.	• Due to the complexity of the architecture, the construction and maintenance costs may be high.
Opportunities	**Threats**
• Provides the ability for easier data processing and development of the data warehouse; • Enables the identification and resolution of data format inconsistencies; • Customization of the data warehouse for different user groups; • Facilitates the creation of trend reports, exception reports, and reports showing actual performance against the organization's goals; • Provides the ability for easy data mining and extracting information from the data.	• Duplication of functions between the data warehouse and the data sources may occur; • Costs increase as functionality grows.

Conclusion: Depending on the required functionality, these three concepts are widely applied in the structuring, storage, and processing of data in a web environment. These architectures can include an unlimited number of databases, which can be distributed across servers at strategic locations worldwide.

1.4. Existing approaches to data storage and management

Bill Inmon is considered the founder of the modern concept of data warehouses, as he first introduced the term "Data Warehouse" in 1990.

After Bill Inmon defined and proposed a methodology for building data warehouses, this methodology became the standard practice for almost six years. In 1996, Ralph Kimball,

a database expert, developed a competing model for the data warehouse. Kimball defines the data warehouse as "a copy of transactional data, specifically structured for querying and analysis" [66], adding that the goal of the data warehouse is "to provide information to support decision-making within a company" [67]. According to Kimball, a data warehouse is a database specifically designed for analysis and decision-making.

Due to the significant increase in data, a new concept called "data lake" emerged in 2010. This term was introduced by James Dixon from the company Pentaho.

Data lake construction strategies can include both SQL and NoSQL approaches to databases, as well as capabilities for online analytical processing (OLAP) and online transaction processing (OLTP).

Table 1.5 summarizes the key differences between the data warehouse and the data lake.

Table 1.5 Differences between the data warehouse and the data lake

Criterion	Data Warehouse	Data Lake
Data	Structured data	Structured, unstructured, semi-structured data
Data Processing	Defined structure before loading Extraction, Transformation, and Loading (ETL) of data into the warehouse	No defined structure; transformation as needed
ETL Process	Extract, Transform, and Load data into the warehouse	Extract and load data into the lake; transform as needed
Costs	High costs	Low costs (in terms of software and hardware)
Technology Maturity	High technology maturity	Technologies in the process of improvement
Users	Business users, data analysts	Data researchers, data analysts

1.5. Conclusions

Based on the analysis of architectures and methods for data structuring, storage, and processing on the Internet, the following conclusions can be drawn:

1. Well-known methods and models for data structuring, storage, and processing, as well as various cloud database architectures, have been analyzed.
2. The classical data warehouse architecture, the architecture with a staging area, and the architecture with a staging area and data marts have been examined to identify differences in their approaches and applications.
3. The specifics of methods, models, and architectures for data structuring, storage, and processing have been described in detail.
4. A SWOT analysis has been performed, evaluating the strengths and weaknesses of different models for data structuring, storage, and processing, helping in the selection of the appropriate architecture.
5. Each model has its own strengths and weaknesses, necessitating a careful choice of the appropriate method for data structuring.
6. The development of web-based information systems requires the integration of intelligent data processing elements into the global network to meet the needs for scalability and adaptability to handle large volumes of data.

7. The application of a hybrid approach to structuring, storing, and processing data minimizes the impact of the weaknesses of individual methods.
8. The choice of the appropriate approach is often influenced by financial and technical factors, with financial factors focused on cost optimization, and technical factors including:
    - Easy and fast development;
    - Scalability potential;
    - Easy maintenance and management;
    - Performance in executing client queries.
9. The wide application of Inmon's and Kimball's models in building solutions for data structuring, storage, and processing leads to their continuous improvement, making them relevant and suitable for the tasks addressed in this dissertation.
10. Through the enhancement of approaches for applying Inmon's and Kimball's models, the necessary long-term perspective for successfully solving the tasks and achieving the goals of this dissertation is achieved.

1.6. Aims and objectives of the dissertation

The aim of this dissertation is the modelling of data structuring, storage, and processing operations on the Internet.

In accordance with the stated aim, the following main objectives are addressed:

1. Modify and investigate methods for structuring, storing, and processing data.
2. Develop and investigate algorithms for structuring, storing, and processing data.
3. Develop models of operations for structuring, storing, and processing data on the Internet.
4. Investigate and compare the proposed models.
5. Identify the factors that characterize information systems as systems for structuring, storing, and processing data on the Internet.
6. Apply the developed models in the design and development of systems for structuring, storing, and processing data on the Internet.

CHAPTER II: METHODOLOGY FOR OBJECTIVE IMPLEMENTATION. CREATION AND PRESENTATION OF MODELS FOR OPERATIONS IN DATA STRUCTURING, STORAGE, AND PROCESSING

Chapter II of the dissertation consists of two parts. The first part presents a methodology for selecting and evaluating models for data structuring, storage, and processing, which includes: choosing an approach for data structuring, storage, and processing, determining an appropriate concept for implementing the assigned tasks, and describing the specific activities to be carried out during the research. The second part discusses the activities related to the preparation and analysis of models for data structuring, storage, and processing on the Internet.

2.1. Methodology for the selection and evaluation of models for data structuring, storage, and processing

The scientific task addressed in this dissertation is related to modeling operations for data structuring, storage, and processing on the Internet. To this end, a hybrid modeling approach has been selected, which uses the fundamental methods for data structuring, storage, and processing, as detailed in Chapter One. This approach enables the combination of dimensional architectures and 3NF architectures, thus covering the operational data store

(ODS), which contains normalized data, as well as dimensional data marts. The selection of data modeling methods supporting scientific research was carried out through a systematic analysis of the main criteria for evaluation and comparison: understandability, interpretability and usability, presence of redundant data, search and query performance, applicability in online data processing systems, flexibility, and complexity.

The generally accepted architectural styles [73][74][76] are the following: Centralized, Independent Data Marts, Federated, Hub-and-Spokes, and Data Mart Bus, as shown in Figure 2.2.

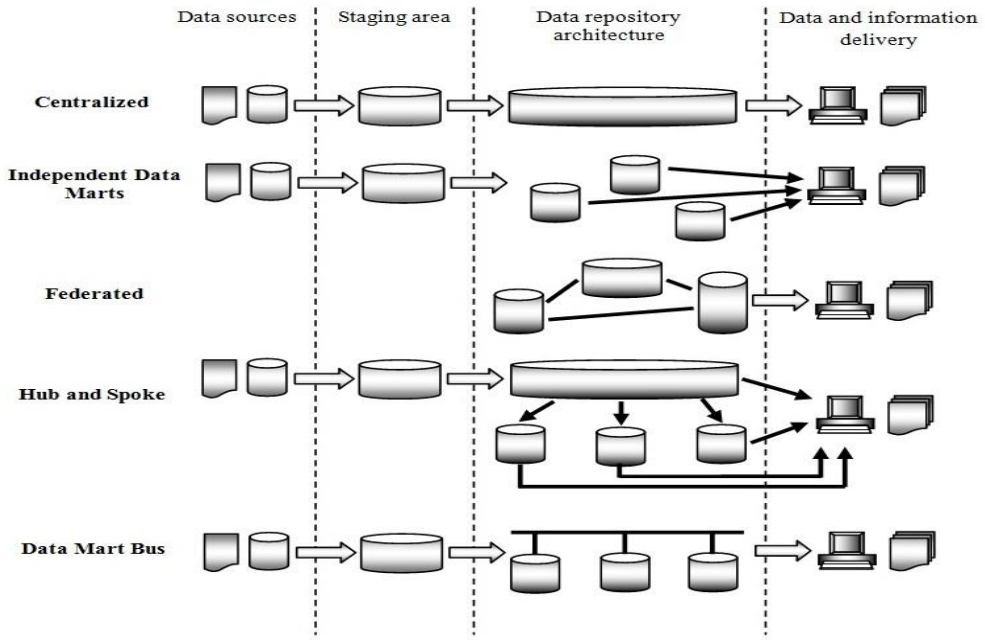


Fig. 2.2 Commonly accepted basic architectural styles of data warehouses

The description of the research conducted and the scientific results obtained is presented for each of the considered models: The Inmon model, the Kimball model, and the Data vault model. The main activities carried out during each stage of the work are based on a combination of analytical and experimental approaches.

Based on a combination of analytical and experimental modeling approaches, the necessary steps in the research methodology of the dissertation have been outlined as follows:

- Review of literary sources detailing various approaches to data structuring, storage, and processing within the domain of cloud technologies;
- Selection of models and data: Examination of their essence, origin, methods of collection, organization, and storage;
- Analysis of available methods and models and initial preparation;
- Visual analysis of the models utilizing software tools such as Microsoft SQL Server, Query Analyzer, and PLATINUM ERwin/ERX;
- Evaluation of the individual impact of each model on the data lifecycle;
- Development of models through the application of selected methods and algorithms for data structuring, storage, and processing;
- Verification and evaluation of the developed models for data structuring, storage, and processing;
- Formulation of conclusions and recommendations for end-users of the data and information management system.

2.2. Creation and presentation of models for data structuring, storage, and processing

The scientific research presented in this dissertation was conducted based on the results of Chapter I, which explores various methods and models for structuring, storing, and processing data on the Internet.

The hybrid model for structuring, storing, and processing distributed data (Fig. 2.3) consists of three levels, grouping containers (data repositories, sources, and users) and processes that cover typical functional groups: extraction, transformation, and loading (ETL), data storage, integration, and data delivery. The included components provide a comprehensive data warehouse (DW) architecture, ensuring secure communications as well as supervisory and analytical functions.

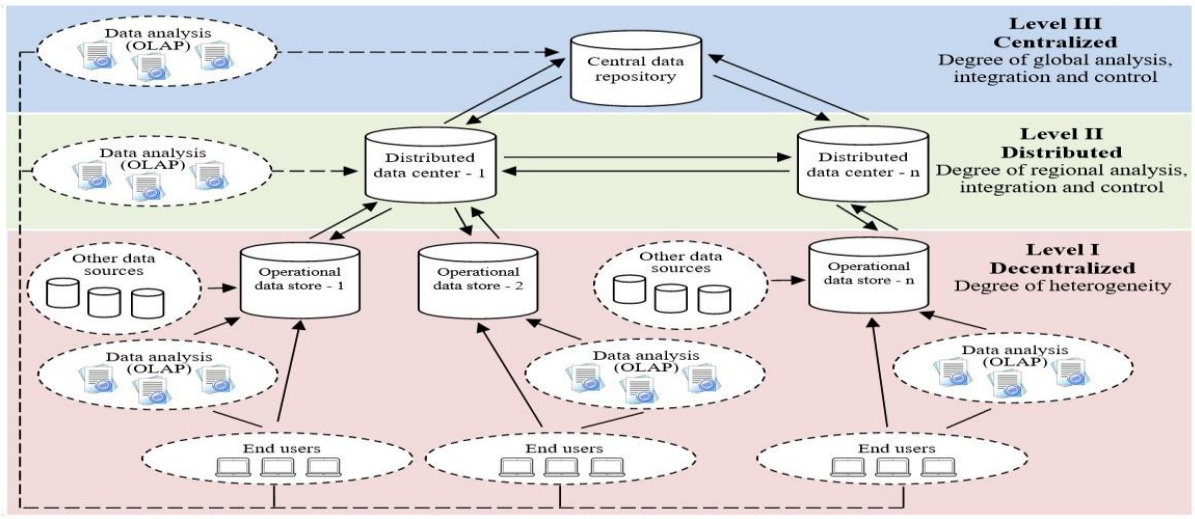
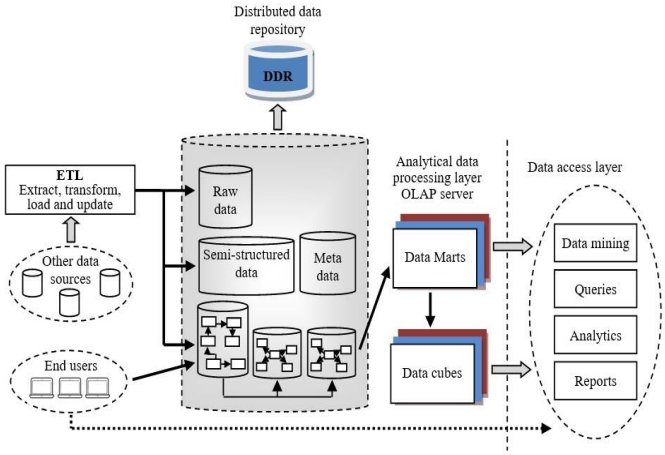


Fig. 2.3 Hybrid model for data structuring, storing and processing

The proposed model offers the opportunity to develop and implement information systems based on current needs and financial constraints. Initially existing structural units for data storage can be integrated, and new ones can be added later, allowing for the system to expand.

Data architecture and data flow architecture are fundamental to data structuring, storing and processing. Data modeling is a basic process when creating a data architecture, which purpose is to determine data arrangement order in each repository so that it can adequately represent business processes. On the other hand, the data flow architecture determines how the individual data repositories are arranged in the data structuring, storage and processing systems, and how the data are transmitted to end users through these repositories.

Level I *- Decentralized (local) structuring, storage, and processing of operational data*

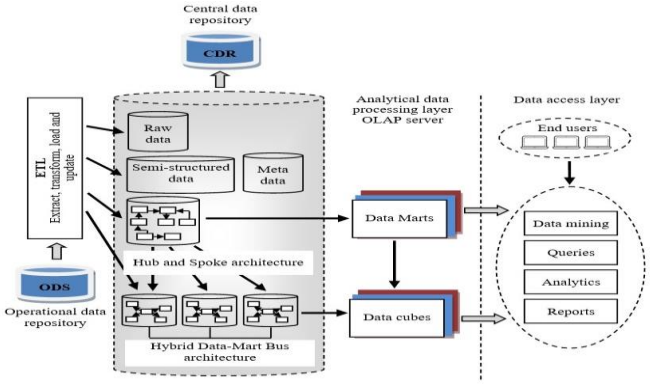


The level includes decentralized operational data repositories [83][84]. The operational data store (Fig. 2.4) functions as a centralized database designed to provide a consolidated view of the most recent data collected from multiple transactional operational reporting systems.

In this repository, data from different sources is aggregated and stored, making it readily available for generating a variety of reports.

Fig. 2.4 Operational data store model

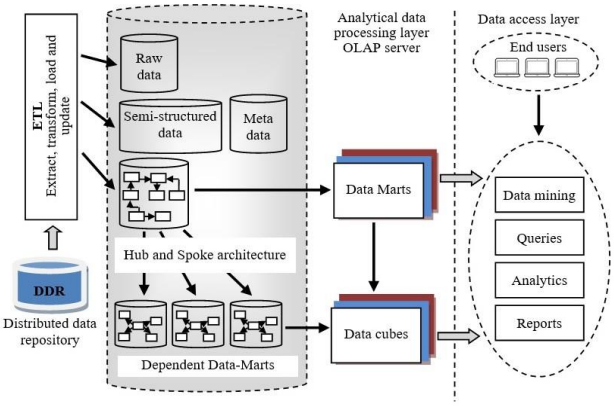
Level II *- Distributed data storage and processing, and access management for data stored in level I*



The distributed data center (Fig. 2.5) is a single node within a framework that utilizes multiple distributed nodes of the same type. These nodes can be managed to enhance performance, security, and optimize network traffic across the global network.

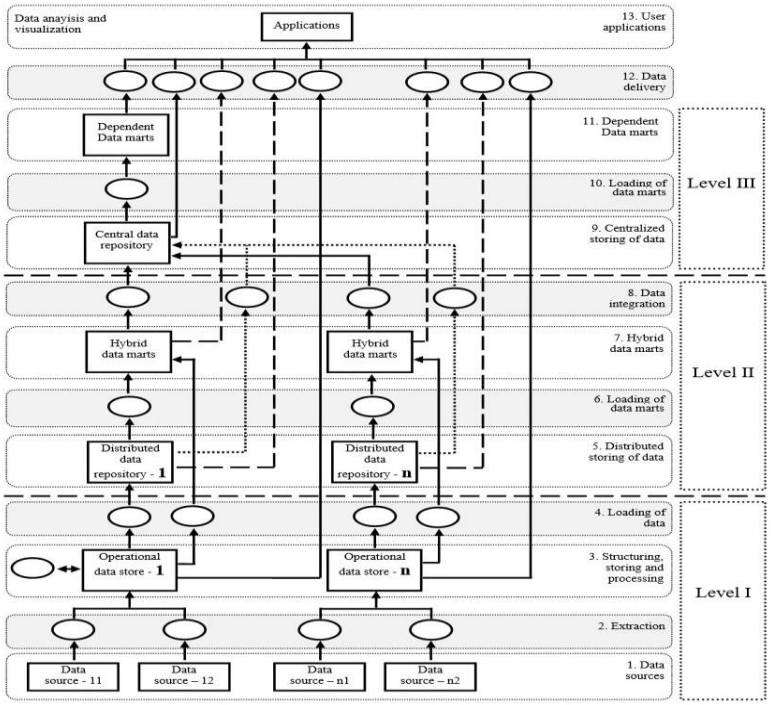
Fig. 2.5 Distributed data center model

Level III *- Centralized data storage and processing, and access management for data located in levels I and II*



The Level represents a central data repository that provides centralized structuring and storage of aggregated data, along with access management to the data located at Levels I and II. In the proposed model, the central data repository (Fig. 2.6) consists of a centralized normalized database that stores data in atomic values, along with dependent data marts [85]. The centralized database, as in Level II, stores data in a normalized form, primarily in third normal form.

Fig. 2.6 Central data repository model



The diagram in Fig. 2.7 presents an architecture for structuring, storing, and processing data, aiming to combine the dominant approaches used to describe such architectures. The main elements in the diagram include processes (abstractions of behavior) and containers, representing components capable of storing, delivering, or consuming data. These may include data marts, data repositories, sources, and others.

Fig. 2.7 Data flow diagram in the hybrid model

2.3. Conclusions

Based on the analysis of the proposed architectures and methods for structuring, storing, and processing data, the following conclusions have been drawn:

1. Traditional information systems based on classical relational databases have several limitations, such as poor adaptability to dynamic changes and difficulty in processing imprecisely formulated or unstructured queries. This drives the development of new intelligent models capable of more flexible and adaptive data structuring and processing.
2. Building intelligent Internet-based systems is a priority for many research teams, as they offer better opportunities for dynamic data structuring and flexible adaptation to new user needs.
3. Current DBMS-based processing solutions are limited and require innovative models that enable more automated and adaptive data management on the Internet.
4. A hybrid model has been developed that combines various data warehouse architectures. This model meets diverse technical and business requirements, offering greater flexibility.
5. The model enables the integration of data from various sources and consists of three main levels:
    - **Decentralized level:** for local storage and processing;
    - **Distributed level:** for data sharing and integration;
    - **Global level:** for centralized data management and aggregation.
6. The architecture of the model ensures a seamless flow and integration of all data entering the warehouse. This enables the creation of a single data source for the entire system, which includes:
    - **Centralized tables** (entity-relationship schema);
    - **Data marts** (star schema) to optimize data access.
7. Phased centralization of data, enabling the use of the benefits of decentralized, distributed, and centralized approaches to storage and processing.
8. The model combines the methods of Inmon and Kimball through three types of data marts: dependent, independent, and hybrid. These provide compactness and flexibility by dividing the data into smaller, subject-oriented sets, enabling easier access for end users.
9. Data marts are created using two approaches:
    - **Dependent data marts:** based on an existing data warehouse, following a top-down approach (Inmon's method);
    - **Independent data marts:** created from internal operational systems or external data, using a bottom-up approach (Kimball's method).
10. The flexibility and efficiency of hybrid data marts are fundamental in building a distributed, geographically structured cloud platform. Such a system provides fast and efficient data access and analysis, enabling the creation of transient or long-term data clusters based on analytical needs.

CHAPTER III: METHODOLOGY FOR CONDUCTING THE RESEARCH. STUDY OF THE FEATURES AND CAPABILITIES OF A "HYBRID MODEL FOR STRUCTURING, STORING, AND PROCESSING DATA ON THE INTERNET"

The third chapter of the dissertation consists of two parts. The first part presents the methodology for conducting the study, which includes a description of the key theoretical concepts for measuring and assessing effectiveness, approaches and methods for analyzing external and internal factors, selecting an appropriate concept for implementing the set tasks, as well as a description of the specific activities to be carried out within the scope of the study. The second part outlines the activities related to the preparation and analysis of models for structuring, storing, and processing data on the Internet.

3.1. Description of the study methodology

In this case, the performance assessment is conducted using two sets of tools. The first set includes studies and tests of SQL-based database management systems (DBMSs), which present various approaches and metrics for evaluating performance. Common evaluation criteria include query execution speed, scalability, reliability, and costs. The second set involves performing a SWOT analysis, which highlights the strengths and weaknesses of the model, as well as the opportunities for mitigating the impact of weaknesses when building hybrid systems for structuring, storing, and processing data. The results of the assessment based on the main performance criteria and the SWOT analysis are presented in section 3.3.

3.2. Preparation for conducting a comparative test and performing a SWOT analysis

To evaluate the performance of SQL Server on typical operations such as read, write, update, and delete, various benchmarking tools, such as YCSB [90] and Sysbench [91], are used. These tools are commonly employed to assess database performance on workloads that simulate real-world applications.

For the purposes of the SWOT analysis, a maximum number of characteristics for strengths, weaknesses, opportunities, and threats are generated within the four categories of the analysis. A total of 53 characteristics are identified and organized into a SWOT matrix, which categorizes them as follows:

- **Strengths** – 16 characteristics;
- **Weaknesses** – 11 characteristics;
- **Opportunities** – 19 characteristics;
- **Threats** – 7 characteristics.

The distribution of features by category is presented in Table 3.5, with the main highlights being:

- Approximately 30% of the generated features represent the strengths of the model;
- About 21% of the features are related to the weaknesses of the model;
- Around 36% of the features reflect the opportunities provided by the model;
- Approximately 13% of the features relate to the threats associated with the implementation, security, and management of information systems built based on this model.

In conclusion, it can be noted that more than half (66%) of the total generated characteristics favor the advantages provided by the model, while the remaining 34% describe its disadvantages. This ratio justifies continuing our research using the SWOT analysis method.

3.3. Results of the Comparative Test and SWOT Analysis

In our case, we used the SQLIO benchmarking tool developed by Microsoft. Alternatives, such as IOMeter, an open-source software with versions for Linux, Solaris, and even NetWare, are also available. However, since SQL Server operates on Windows, SQLIO is the more appropriate choice for our case.

After editing the configuration file, we run SQLIO once to create **testfile.dat**. This is done via the command line with the following command:

sqlio -kW -s5 -fsequential -o4 -b64 -Fparam.txt

Test results:

To test device D, we create a series of tests in a batch file (test.cmd) and let them run. The goal is to determine the maximum number of I/O operations that device D can handle.

Random write test

The values for **input/output operations per second (IOPS)** and **megabytes per second (MB/sec)** are summarized in Table 3.1.

Table 3.1 Random write test results

Pending input/output operations	Input/output operations per second	Megabytes per second
1	100.05	6.25
2	98.85	6.17
4	99.92	6.24

Random Reading Test

The input/output operations per second (IOPS) and megabytes per second (MB/sec) test results are summarized in Table 3.2.

Table 3.2 Random read test results

Pending input/output operations	Input/output operations per second	Megabytes per second
1	92.24	5.76
2	197.54	12.34
4	217.85	13.61
8	262.88	16.43
16	313.39	19.58
32	353.11	22.06
64	393.25	24.57
128	384.83	24.05
256	390.36	24.39
512	385.95	24.12
1024	395.28	24.70

Sequential reading test

The input/output operations per second (IOPS) and megabytes per second (MB/sec) test results are summarized in Table 3.3.

Table 3.3 Sequential read test results

Pending input/output operations	Input/output operations per second	Megabytes per second
1	2227.39	139.21
2	2233.41	139.58
4	2190.09	136.88
8	2217.39	138.58
16	2236.75	139.79
32	2262.47	141.40
64	2300.28	143.76
128	2305.87	144.11
256	2286.90	142.93

Pending input/output operations	Input/output operations per second	Megabytes per second
512	2288.37	143.02
1024	2295.21	143.45

Sequential write test

The input/output operations per second (IOPS) and megabytes per second (MB/sec) test results are summarized in Table 3.4.

Table 3.4 Sequential read test results

Pending input/output operations	Input/output operations per second	Megabytes per second
1	6813.03	53.22
2	6339.24	49.52
4	10323.31	80.65
8	7739.09	60.46
16	7645.11	59.72
32	8604.67	67.22
64	9164.68	71.59
128	9044.44	70.65
256	9409.32	73.51
512	9365.35	73.16
1024	9455.53	73.87

SQL Server demonstrates impressive performance in standard data manipulation operations, particularly when optimization strategies are applied and sufficient hardware resources are utilized. Effective resource management and the implementation of batch operations are key factors in system efficiency, especially for high-performance applications. Benchmarking analyses underscore the importance of careful planning and tuning to achieve optimal performance of SQL Server in real-world scenarios.

Based on the basic matrix, a detailed SWOT matrix has been created, which groups the factors of the model into four aspects. Strengths and capabilities are primarily identified as factors that have a positive impact on the structuring, storage, and processing of data. Weaknesses and threats are mainly determined by factors arising from the operational characteristics and complexity of the proposed system. Table 3.5 presents a summary SWOT matrix of the hybrid model for structuring, storing, and processing distributed data on the Internet.

Table 3.5 Summary SWOT matrix of the Hybrid model

Strengths	**Weaknesses**
Level I - Decentralized (local) structuring, storing, and processing of operational data.	
Local structuring, storage, and processing of operational data.	As the number of nodes increases, it becomes progressively more difficult to ensure that users receive the same view of the database.
Easy networking of systems containing local databases and adding additional nodes.	Data management, related to the simultaneity and integrity of the data, is not fully guaranteed.
Low costs during initial construction and implementation.	Data redundancy is possible, as they are stored in multiple locations.
Load resistance, as data processing can be performed locally.	Complex software is required to manage the communication between the nodes.

Strengths	Weaknesses
Fault tolerance – in the event of a node failure, the rest of the system can continue to operate. This depends on the software controlling the communication between nodes.	Challenges in finding specialists with the necessary experience in managing and maintaining distributed systems.
Structuring, storing, and processing data by end users.	
Level II - Distributed storing and processing of data, and access management to data stored in Level I.	
Local workload during data processing and access provision.	Difficulty in ensuring a consistent view of the database as the number of nodes increases.
Planning of replication and duplication processes depending on the system load.	Data management, related to the simultaneity and integrity of the data, is not fully guaranteed.
In some cases, when the users and their data are in the same region, the advantages of centralized data storage and processing may apply.	Data redundancy is possible, as they are stored in multiple locations.
Fault tolerance – in the event of a node failure, the rest of the system can continue to operate. This depends on the software controlling the communication between nodes.	Ensuring data security is challenging due to the decentralized nature and complexity of the infrastructure connecting all nodes.
Data can be grouped and stored in thematically oriented areas based on a specific criterion (such as departments within an organization), thereby organizing hierarchical access.	Complex software is required to manage communication between the nodes.
	Inconsistencies may arise between replicated data if some of it is changed.
	Challenges in finding specialists with the necessary experience in managing and maintaining distributed systems.
Level III - Centralized storage and processing of data, with access management to the data located in Levels I and II.	
Data integrity is guaranteed, and data redundancy is minimized due to the normalized nature of the data.	High dependence on network connectivity. The slower the internet connection, the longer it will take to access the database.
Higher data security, as the data is stored in a single location.	Potential for failures when a large number of users work simultaneously.
Data stored in a single repository are easier to manipulate, such as updating, reorganizing, or analyzing.	Minimal or no data redundancy is a prerequisite for partial data loss.
The data can be accessed simultaneously from multiple locations.	Access by multiple users is restricted to a single data set, which can significantly reduce the overall efficiency of the system.
Data updates are instantly available to all users.	
Low costs for labor, energy, and maintenance.	
Simplified design makes it easier for end users to operate.	

Opportunities	Threats
Level I - Decentralized (local) structuring, storing, and processing of operational data.	
Provides easy access to operational data and processing of simplified queries.	Data integration and consolidation from multiple sources can become complex, especially when comparing data from two different systems.

Opportunities	Threats
Provides information for operational and tactical decision-making based on real-time or near-real-time data, considering time zone differences for immediate use.	Difficulties in ensuring the quality of data transformation and integration. The degree and type of data transformation depend on the frequency of data delivery. The shorter the delivery time, the lower the quality of transformation and integration. One solution to this issue is the use of the Extract, Transform, Move, Load (ETML) tool, which efficiently handles large volumes of updates and supports a wide range of data structures and formats.
Integration of data from both new and existing systems through centralized data storage, which enhances reporting capabilities. This approach goes beyond the limited reporting offered by individual system sources, enabling the creation of more comprehensive operational reports.	The growing volume of data results in higher management costs.
Providing a comprehensive view and monitoring of the stored data, which facilitates the identification and diagnosis of issues within the information system.	
Acting as an intermediary station before data is transferred to distributed data centers, while also performing operations to enhance data quality.	
Providing end users with the ability to structure, store, and process data according to their needs.	
Level II - Distributed storing and processing of data, and access management to data stored in Level I.	
Integrating data from decentralized operational data stores (ODS) and implementing centralized data storage based on a regional model. This enables the generation of reports and summary analyses for each region.	As the number of tables grows, the E-R model and query execution may become more complex, involving additional tables and unions.
Easy access to the summarized historical data retrieved from the ODSs located at Level I.	The necessity for experts in data modeling and business processes.
Providing information for analysis and decision-making based on historical data.	The data stored in the data marts (DM) are not fully integrated, and data redundancy may occur.
Reduction of data redundancy through normalization, avoiding anomalies during updates.	Adding columns to the fact table can degrade performance as its size increases, making it harder to adapt the dimensional model to changing business requirements.
Ability to combine data from the data repository with other data sources.	
Improving performance through the use of denormalized, object-oriented hybrid dimensional data marts (HDM).	
The ability for end users to utilize specialized data analysis software.	
Providing a comprehensive view and monitoring of the stored data, which facilitates the identification and diagnosis of issues within the information system.	

Opportunities	Threats
Performing the functions of an intermediate stage before the data is transferred to the central data repository, along with operations to enhance their quality.	
Level III - Centralized storage and processing of data, with access management to the data located in Levels I and II.	
Integration of data from the distributed regional data centers, enabling the generation of various reports and global analyses.	As the number of tables grows, the E-R model and query execution can become more complex due to the inclusion of additional tables and unions.
Easy access to the summarized historical data retrieved from the distributed data centers (DDC) located at Level II.	The need for experts in data modeling and business processes.
Providing information for analysis and decision-making based on historical data.	The data stored in the DM are not fully integrated, and data redundancy may occur.
Data redundancy is reduced through normalization, which helps avoid anomalies during updates.	Adding columns to the fact table can degrade performance as its size increases. This also makes it challenging to adapt the dimensional model to changing business requirements.
The data repository is the only source of data that guarantees its integrity and consistency.	
Increasing the performance by using denormalized object-oriented dimensional dependent data marts (DDM).	
The possibility for end users to utilize specialized data analysis software.	
Providing a comprehensive view and monitoring of the stored data, which facilitates the identification and diagnosis of issues in the operation of the information system.	
Performing the functions of a global data repository, as well as carrying out operations to enhance data quality.	

- Strengths

The hybrid model provides several key advantages, including the ability to be built incrementally, low initial implementation costs, load resilience, and efficient management of data storage, processing, and access. It also offers strong data security and fast performance. However, these benefits come with the need for skilled professionals to manage and maintain the smooth functioning of all components of the data warehouse. Additionally, the model's capacity to generate a variety of reports is another important advantage.

- Weaknesses

The main weaknesses of the model include challenges in ensuring concurrency and data integrity at the distributed level. As the number of nodes increases, maintaining a consistent view of the data for users becomes more difficult, and data redundancy may arise from simultaneous storage in different locations. Another drawback is the requirement for complex software to manage communication between distributed nodes and handle data modeling, making it harder to find specialists with the necessary expertise. Additionally, reporting can become more challenging as the number of tables grows and business rules evolve, especially when dealing with complex queries.

- Opportunities

The gradual centralization of storage offers opportunities to extract the necessary information for generating more detailed reports and making operational and tactical decisions based on historical, current, or near-real-time data [95]. Users have the flexibility to structure, store, and process data as needed and utilize specialized software for data analysis. By combining data from the repository and other sources through hybrid data marts, rapid integration of multiple databases is enabled, merging the advantages of both independent and dependent data marts [96]. This approach also provides a comprehensive view and monitoring of the stored data, which facilitates the identification and diagnosis of issues within the information system [97].

- Threats

Data integration and consolidation can become challenging when matching data from different systems. A lack of sufficient time for extraction and loading may negatively impact data quality, transformation, and integration. Additionally, the complexity of the model, as the number of tables increases, can complicate query execution and report generation. This complexity also hinders the adaptation of the model to changing business requirements and increases management and maintenance costs, as it necessitates experts in data modeling and business processes. The high cost of management and maintenance is a significant threat, but it can be mitigated by either training in-house employees or hiring specialists for this purpose.

3.5. Conclusions

As a result of the SQL Server performance test and the SWOT analysis of the "Hybrid Model for Structuring, Storing, and Processing Distributed Data" the following conclusions and insights can be drawn:

1. SQL Server is capable of processing a large volume of queries when operating under optimized hardware settings.
2. Using batch updates significantly improves runtime and reduces resource consumption.
3. Group operations, such as data updates and deletions, enable strategic task separation, resulting in improved performance.
4. Proper SQL Server configuration is crucial for achieving high performance in workload-intensive applications.
5. The model integrates various storage architectures that meet both technical and business requirements. This integration offers a more flexible and adaptable data management environment.
6. Advanced methods for structuring and processing data are proposed, with a focus on the efficient sharing and integration of data from diverse sources. This approach addresses the challenges of unifying heterogeneous information flows.
7. Each architecture at the different levels of the model has its own unique strengths and weaknesses. Therefore, selecting the appropriate modeling method for the specific system is essential.
8. Due to its adaptability, the model allows trade-offs between different architectures and capabilities, while also enabling phased implementation.
9. The hybrid model offers fast deployment, easy maintenance, scalability, and high performance for customer requests. Its easy development and flexible expansion ensure the system can adapt to changing requirements and workloads.
10. By integrating existing structures and adding new ones based on needs and resources, the model achieves better manageability and cost control.

CHAPTER IV: SOFTWARE SOLUTION FOR DATA STRUCTURING, STORAGE, AND PROCESSING OPERATIONS OVER THE INTERNET

This chapter presents the implementation of a data and information management system based on the hybrid model for structuring, storing, and processing distributed data over the Internet, as described in Chapter II. It defines the system requirements, scope, technology and tool selection, architecture, workflow organization, and the results obtained.

4.1. Good data management practices

Data management involves actions for structuring, storing, and processing data used or generated by various processes, such as research work [98], business relations, or data generated by end devices. A common practice is to retain data for at least 10 years [99][100], which fully meets the requirements for traceability, long-term interpretability, and sustainable archiving. Defining basic principles can contribute to good data management by improving discoverability, accessibility, interoperability, and reusability.

4.2. Data structure and formats

The structure of generated data largely depends on the software used. Data formats can be binary, commonly accepted ASCII or XML code, or other specific formats defined by the applications in use. The need for structuring data generated by users and storing it in a specific schema, along with the requirements for subsequent collaborative use of this data by other users, are key topics in this research.

4.3. Workflows data management

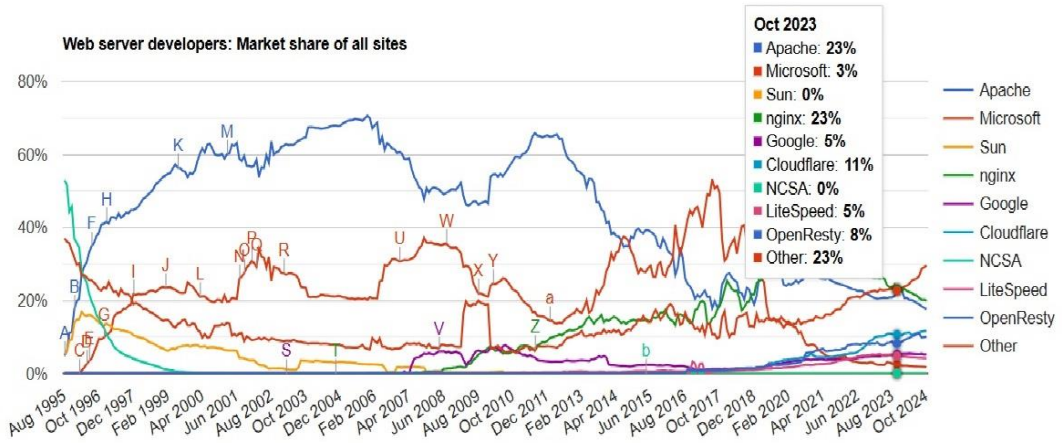
Workgroups, in their activities, create workflows that generate and process data specific to their respective processes. This diversity makes it challenging to establish unified data management. Regulating access to stored data is a fundamental aspect of building any local or web-based information system, ensuring reliability and protection against unauthorized access.

4.4. System requirements

The proposed methodology for data and information management necessitates the use of appropriate tools for conducting scientific experiments. The system is designed to support scientific and experimental work, enabling users to define their preferred data storage structure and perform various operations, such as data input, editing, and sharing. These functionalities will greatly simplify access to databases for a wide range of users.

4.5. Selection of tools and technologies for implementation

Figure 4.1 illustrates the distribution of the main web platforms used by website developers in recent years.



Developer	September 2024	Percent	October 2024	Percent	Change
nginx	225,640,032	20.16%	226,359,600	20.01%	-0.15
Apache	201,390,151	18.00%	199,150,231	17.61%	-0.39
Cloudflare	130,093,325	11.63%	131,624,333	11.64%	0.01
OpenResty	111,723,893	9.98%	113,940,338	10.07%	0.09

Fig. 4.1 Market share distribution of active sites according to Netcraft

The graph in Fig. 4.1 shows that over time the popularity of NGINX decreases and gradually approaches that of Apache HTTP Server.

Table 4.1 presents the main characteristics of the two web platforms. Summarizing and organizing these characteristics creates a structured information base for decision-making and aids in selecting the most suitable web server.

Table 4.1 Summary of the main features of Apache and NGINX

Apache	NGINX
Its primary function is to serve as a web server.	It has two functions: as a web server and a reverse proxy server.
It cannot handle multiple requests simultaneously under heavy traffic.	It can handle multiple requests simultaneously, but with limited resources.
It uses a multi-threaded approach to process requests.	It uses an event-driven approach for request processing.
It supports the dynamic loading and unloading of functional modules, enhancing flexibility.	Functional modules cannot be loaded dynamically. They must be integrated into the main software.
It processes dynamic content directly on the web server.	By default, it cannot process dynamic content.

Apache HTTP Server was selected to implement a software solution for structuring, storing, and processing data on the Internet.

4.6. Data and information management system

The system is based on the methods for structuring, storing, and processing data on the Internet, as discussed in Chapter one, as well as on a hybrid model for these processes, presented in Chapter two.

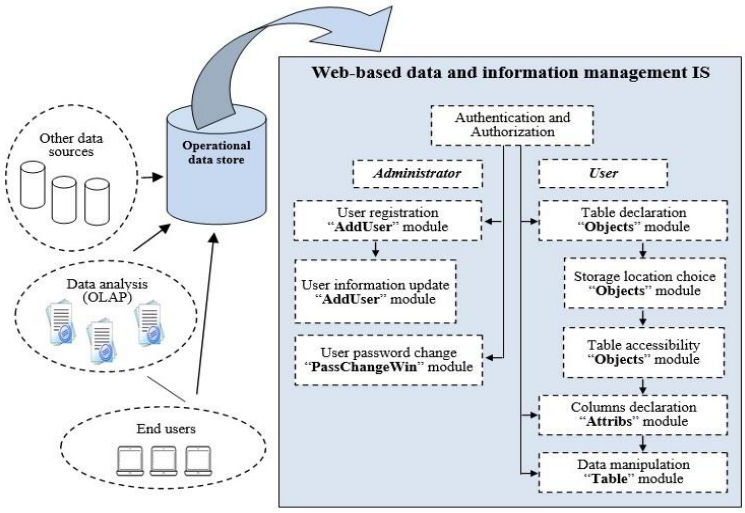


Figure 4.3 presents a summary implementation view of the system within an operational data store. This data store is part of Level I of the hybrid model for structuring, storing, and processing data on the Internet, providing decentralized (local) structuring, storage, and processing of operational data.

Fig. 4.3 Summary implementation view of the system

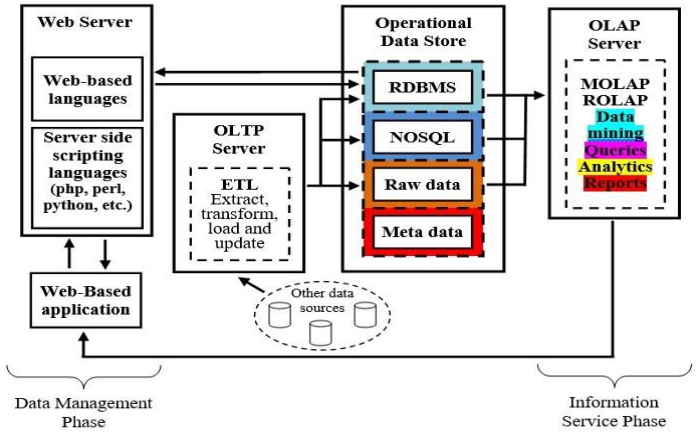


Figure 4.4 illustrates the software and hardware requirements of the system components and their interactions. The hardware and software components selected for the implementation process were based on three criteria: cost, availability, and ease of open-source programming.

Fig. 4.4 Block diagram of the system

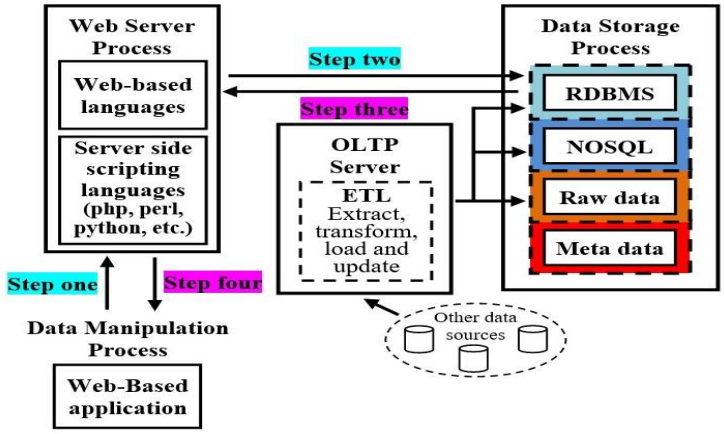


Figure 4.5 illustrates the processes, steps, and procedures performed by users in structuring, storing, and manipulating data within the system.

Fig. 4.5 Procedure steps for data processing phase

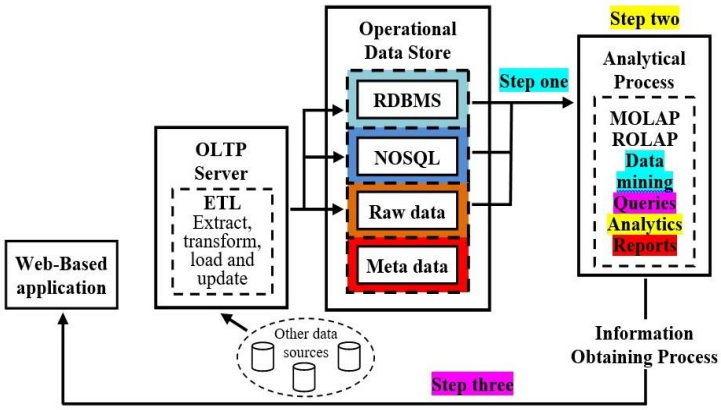


Figure 4.6 provides an overview of all the steps and procedures used to deliver the information service to users.

Fig. 4.6 Procedure steps for the information service phase

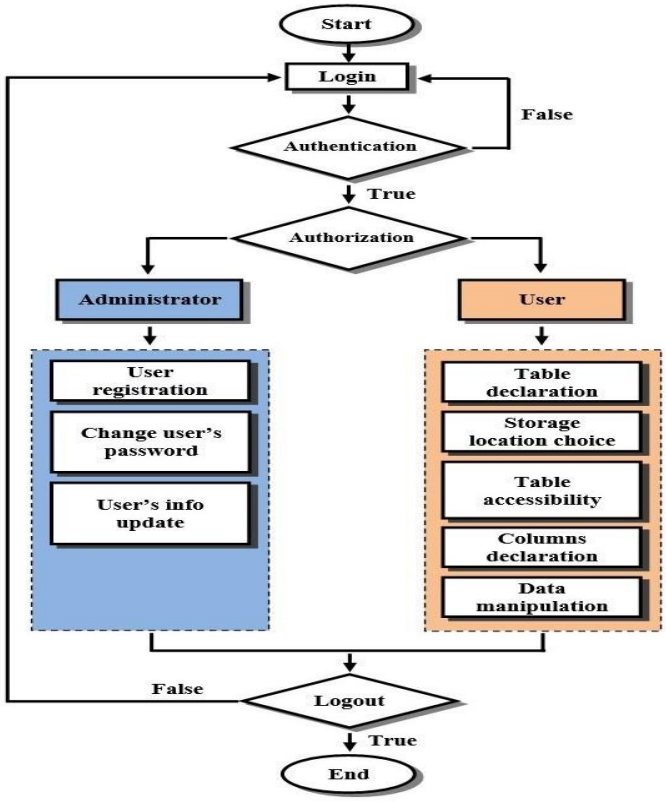


Figure 4.7 outlines the sequence of operations and system modules, which are divided into two main parts: administrative and user sections:

- The administrative part allows for adding users to the database, changing user passwords, and updating user information;
- The user part allows the following operations: declaring a table, selecting a storage location, defining table access, declaring columns, and manipulating data.

Fig. 4.7 Block diagram of the system operation algorithm

The main graphical user interface (Fig. 4.8), displayed after identification, includes a drop-down list with options to declare a new table or select an existing one for the user to work with.

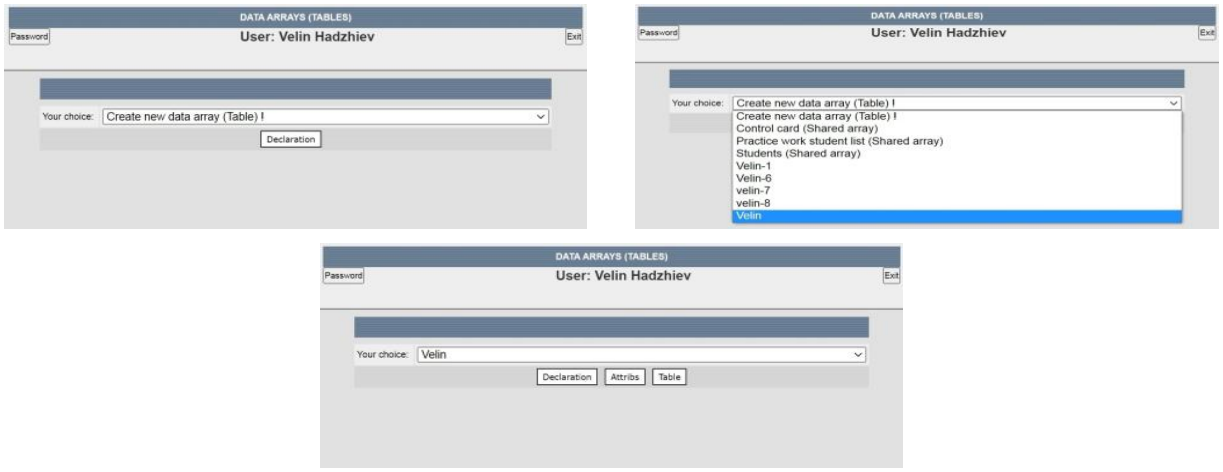


Fig. 4.8 System main interface

Through the "New Table" functional module, the user can declare a new table or edit the parameters of an existing one (Fig. 4.9).

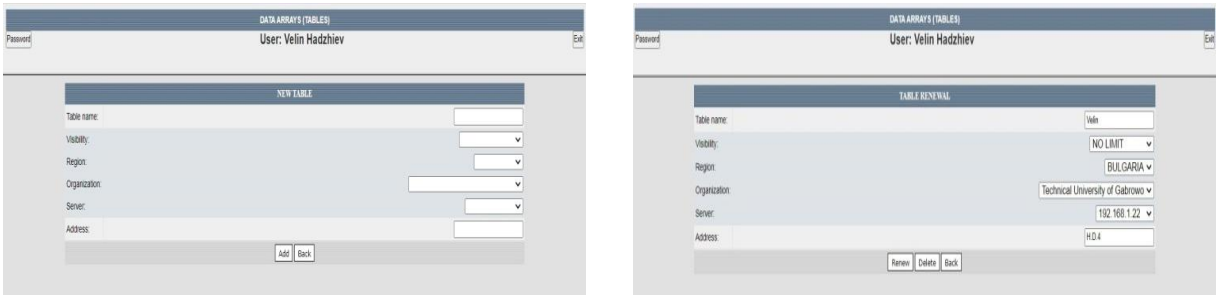


Fig. 4.9 Table parameters

Through the "Table Structure" functional module (Fig. 4.10), the user can either add new columns or edit the parameters of existing ones.

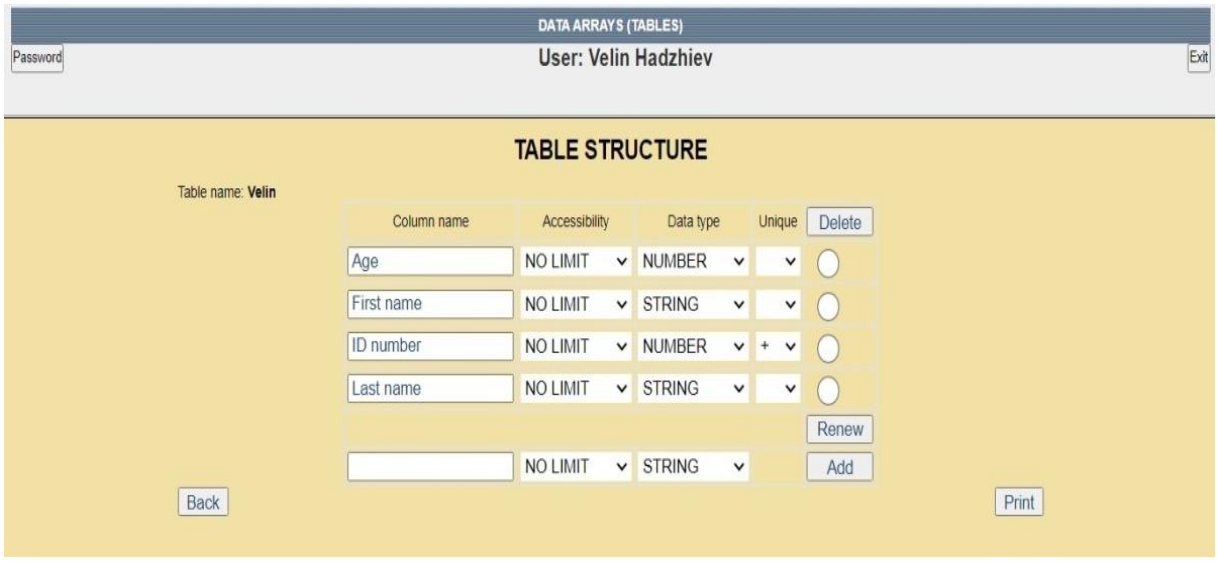


Fig. 4.10 Adding or editing columns

Through the "Table Content" functional module (Fig. 4.11), the user can enter, delete, or edit data in a table.

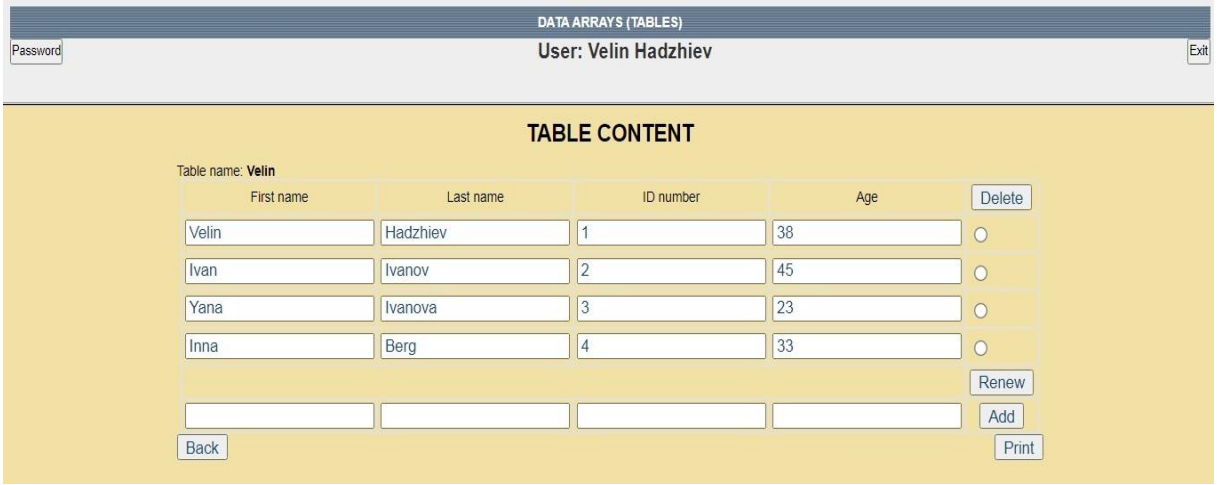


Fig. 4.11 Data manipulation

4.7. Advantages and disadvantages

The SWOT matrix presented in Table 4.3 aims to categorize the factors of the system into four groups: strengths, weaknesses, opportunities, and threats. Classifying these factors into separate groups provides detailed information about the system's advantages and disadvantages. This information can be used to mitigate the impact of weaknesses and serve as a guideline for future work to enhance the system's efficiency.

Table 4.3 SWOT matrix of the system

Strengths	Weakness
A user-friendly approach to structuring, storing, and processing data.	As the number of tables increases, the complexity of the model, implementation, and maintenance also increases.
Structuring and storing an unlimited volumes of data online.	High dependency on network connection: system access time is influenced by internet connection speed.
Storing periodically updated data in a separate space, distinct from the one used for configuration information.	The system may experience failure when a large number of users work simultaneously.
Data backup in the event of a technical failure.	

Opportunities	Threats
The possibility of networking separate systems with local databases and adding additional nodes.	System performance may degrade as the volume of data increases.
Placing frequently accessed data closer to users to reduce web traffic.	
The ability to access data simultaneously from multiple users.	
The ability to provide data from various sources, including databases, directories, email, etc.	

4.8. Conclusions

In conclusion, the developed method for performing basic data structuring, storage, and processing operations, implemented in a web-based data processing and management system, possesses the following characteristics and capabilities:

1. Flexible management of tables and attributes, offering a high level of personalization and control over data.
2. Create an unlimited number of tables, set visibility, define and configure table attributes, and input data into tables based on user rights.
3. Management of both individual and collaborative workflows, regardless of the application domain, making it versatile for a wide range of users.
4. Easily adaptable to various scientific and work processes without altering its core structure or functionality, making it suitable for rapid implementation in new applications and domains.
5. The system is designed to work with relational databases, and its initial implementation and maintenance require specialized expertise, leading to higher costs in the project's early phase.
6. Increased complexity of the database model with a rise in the number of tables, and reduced system efficiency as data volume grows.
7. The presented system offers a high degree of personalization, versatility, and adaptability to various workflows. To address the challenges of cost and complexity in managing large volumes of data, the system can be implemented as a cloud service (DBaaS) or utilize NoSQL databases. These solutions can significantly enhance scalability and efficiency as data volumes and application variety grow.

OVERALL CONCLUSIONS/FINAL REMARKS

As a result of the analysis of methods and models for structuring, storing, and processing data on the Internet, the following conclusions and observations can be made:

1. Traditional methods of structuring, storing, and processing data often fail to easily adapt to changing user needs or the dynamic requirements of the internet environment. Their limited ability to address imprecisely defined tasks further restricts flexibility. As a result, modern data management models on the internet are increasingly being utilized in intelligent information systems (IIS) to address these challenges.
2. In most enterprise DBMSs, queries follow a strictly fixed syntax and semantics, which complicates the process of structuring and accessing data. The limitations of these methods often result in difficulties when processing queries written in natural language. Therefore, innovative data storage and processing models are needed to facilitate working with diverse language structures.
3. One effective method for structuring data is to create systems in which the data structure is described in a way that is easily understood by all users. This approach facilitates the design and construction of databases, even by individuals without specialized technical knowledge. These models significantly enhance the efficiency of data management systems on the internet.
4. To improve data storage and processing methods, modern DBMSs are developing technologies that enable data queries and analysis to be tailored to the human way of thinking and expression. This includes the use of artificial intelligence and machine learning, which allow models to store and analyze data based on the context and needs of the user.

5. One of the most essential methods of processing data on the internet is the creation of models that can extract information and generate answers based on accumulated knowledge. These models are designed to handle analytical and predictive tasks using experience and provide relevant answers in real-time.
6. Remote data access capabilities require processing methods that offer flexibility and scalability. Integrating information systems based on these methods into the internet network enables users to structure and store data from various sources, tailored to their specific needs.
7. As technology advances, it becomes increasingly important to clearly define and understand key terms such as "structuring", "storage", and "processing" of data. Clarifying these concepts is essential for creating effective data management models that meet the dynamic demands of the internet environment and ensure reliable, optimized information management.

In line with the objectives outlined in the dissertation, various methods and models for data processing have been developed, researched, and tested. As a result, the following conclusions and observations can be drawn:

1. Various methods for structuring, storing, and processing data are examined, including both traditional and hybrid cloud database architectures.
2. An analysis of each method and architecture is conducted, highlighting their strengths and weaknesses. This provides detailed insights for selecting an approach to designing information systems that accommodates necessary compromises.
3. Existing approaches to data storage and management, such as data warehouses and data lakes, are reviewed, with a description of their main characteristics and functions.
4. A hybrid model is proposed that combines different data warehouse architectures, offering flexibility and efficiency in managing data from various sources. The model incorporates three organizational levels: decentralized, distributed, and global, enabling a gradual centralization of data.
5. A reference model [73] is proposed, incorporating well-known architectural styles and providing options for adaptation to meet specific goals in structuring, storing, and processing data. The model's architecture can be customized by adding or removing layers, nodes, and components.
6. A SWOT analysis of the proposed model has been conducted, highlighting that the hybrid approach provides solutions suitable for various business needs. However, each architecture has its own limitations, requiring compromises when designing information systems. The model's flexibility allows for gradual implementation and scaling as needed.
7. An intuitive web-based application for structuring and storing data has been developed, making it accessible even to non-specialists.

Chapter 4 of the dissertation presents a straightforward approach to data management through a web-based information system that enables the structuring, storing, and processing of user data on the internet. The data structure is described in a way that is accessible to all users, facilitating the design and construction of databases even by non-experts. The system also supports data sharing among users, promoting collaborative work and the analysis of shared data.

Users can create tables, set their visibility for other users, define attributes, and enter data into both personal and shared tables. The system is suitable for managing individual workflows as well as facilitating collaborative work among multiple users, regardless of the application area. It can be easily adapted to various scientific or work processes that generate or process data, without the need to modify its structure or functionality. Through these features, the system effectively realizes the concept of making databases accessible to a wide range of users.

CONTRIBUTIONS OF THE DISSERTATION

As a result of the research conducted for this dissertation, the key contributions can be summarized as follows:

Scientific and practical contributions

1. A comprehensive review of contemporary literature in the field of data modeling, structuring, storage, and processing on the internet has been carried out.
2. A comprehensive analysis of architectures for data structuring, storage, and processing in cloud environments has been conducted. This analysis serves as the foundation for developing sustainable and scalable systems that meet the requirements for database accessibility for a wide range of users.
3. A methodology has been developed for selecting and evaluating models for data structuring, storage, and processing, offering a systematic approach to adapting these models to specific requirements for data structuring, storage and processing.
4. Models for data structuring, storage, and processing were evaluated and analyzed, with a focus on applying the developed methodology for their assessment. Software tools were utilized to provide an objective evaluation of critical aspects such as efficiency, sustainability, and scalability of the models.
5. A hybrid model has been created that combines key functionalities of the selected models, addressing specific requirements for data structuring, storage, and processing, while ensuring database accessibility for a wide range of users. The developed data flow diagram demonstrates the model's effectiveness in various practical scenarios.
6. A detailed SWOT analysis of the hybrid model has been conducted, confirming its potential for integration into real systems and highlighting its flexibility and resilience across various solutions.

Practical contributions

1. A method for optimizing data operations has been developed, integrating best practices and proven techniques for data structuring, storage, and processing. Its applicability has been demonstrated through simulations and tests in real conditions.
2. The effectiveness of the proposed hybrid model has been validated through empirical testing, which includes assessments of performance, resilience, and scalability.
3. Based on the method for optimizing data operations, a web-based system for data structuring, storage, and processing has been developed, offering access to databases for a wide range of users. Tests conducted in real-world scenarios confirm its practical effectiveness.

LIST OF PUBLICATIONS RELATED TO THE DISSERTATION

[1] Hadzhiev, V., Rashidov, A., "Overview and analysis of methods and models for data structuring, storage, and processing on the Internet", Proceedings of the International Scientific Conference Automation and Informatics '19, Sofia, 2019, Volume 1, pp. 215-218, ISSN 1313-1850.

[2] Hadzhiev, V., Rashidov, A., "Overwiew and analysis of architectures for data structuring, storage, and processing in the cloud", Proceedings of the International Scientific Conference Automation and Informatics '19, Sofia, 2019, Volume 1, pp. 219-222, ISSN 1313-1850.

[3] Hadzhiev, V., Rashidov, A., "Overview and analysis of methods and models for data structuring, storage, and processing on the Internet", Automation and Informatics Journal, 2019, No. 2, pp. 27-32, ISSN 0861-7562.

[4] Hadzhiev, V., Rashidov, A., "Overwiew and analysis of architectures for data structuring, storage, and processing in the cloud", Automation and Informatics Journal, 2019, No. 3, pp. 12-16, ISSN 0861-7562.

[5] V. Hadzhiev, A. Rashidov, "Overview and analysis of methods and models for data structuring, storage, and processing on the Internet", 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 2019, pp. 492-495, doi: 10.23919/ELECO47770.2019.8990416 - Scientific publication - Conference paper presented abroad.

[6] Velin Hadzhiev and Aldeniz Rashidov, "A Hybrid Model for Structuring, Storing, and Processing Distributed Data on the Internet", 2021 International Conference on Automatics and Informatics (ICAI), 2021, pp. 82-85, doi: 10.1109/ICAI52893.2021.9639240 - Scientific publication - Conference paper presented in Bulgaria.

[7] Velin Hadzhiev and Aldeniz Rashidov, "An Organization of the Storage and Data Flow in a Hybrid Model for Structuring, Storing, and Processing Distributed Data on the Internet", 2021 13th International Conference on Electrical and Electronics Engineering (ELECO), 2021, pp. 589-593, doi: 10.23919/ELECO54474.2021.9677648 - Scientific publication - Conference paper presented abroad.

[8] Velin Hadzhiev, "SWOT Analysis of a Hybrid Model for Structuring, Storing, and Processing Distributed Data on the Internet", 2021 13th International Conference on Electrical and Electronics Engineering (ELECO), 2021, pp. 585-588, doi: 10.23919/ELECO54474.2021.9677789 - Scientific publication - Conference paper presented abroad.

[9] Velin Hadzhiev, Aldeniz Rashidov, "Implementation of a Data and Information Management System Based on a Hybrid Model for Structuring, Storing, and Processing Distributed Data on the Internet", 13th International Conference on Computing Communication and Networking Technologies, Online (Technical Assistance - IIT Mandi), India, October 3-5, 2022, doi: 10.1109/ICCCNT54827.2022.9984384 - Scientific publication - Conference paper presented abroad.

CITATIONS OF PUBLICATIONS RELATED TO THE DISSERTATION

Citations from the list of publications related to the dissertation have been identified as follows:

1. **A citation of work #7** in a journal with an impact factor of 7.94 for 2023-2024.

Cited article: Velin Hadzhiev and Aldeniz Rashidov, An Organization of the Storage and Data Flow in a Hybrid Model for Structuring, Storing and Processing Distributed Data on the Internet, 2021, 13th International Conference on Electrical and Electronics Engineering ELECO, 2021, pp. 589-593, doi: 10.23919/ELECO54474.2021.9677648 - Scientific publication - Conference paper presented abroad.

Journal: Computers & Industrial Engineering, Volume 190, 2024, 110082, ISSN 0360-8352

Citing author: Samuel Harno, Hing Kai Chan, Min Guo

Title of the article: Enhancing value creation of operational management for small to medium manufacturer: A conceptual data-driven analytical system

Link: https://doi.org/10.1016/j.cie.2024.110082

(https://www.sciencedirect.com/science/article/pii/S0360835224002031)

Impact factor: 8.78

Link for the impact factor: https://www.resurchify.com/impact/details/18164

2. **Two citations of work #8** in journals with impact factors of 0.71 and 1.03 for 2023-2024.

Cited article: Velin Hadzhiev, SWOT Analysis of a Hybrid Model for Structuring, Storing and Processing Distributed Data on the Internet, 2021, 13th International Conference on Electrical and Electronics Engineering ELECO, 2021, pp. 585-588, doi: 10.23919/ELECO54474.2021.9677789 - Scientific publication - Conference paper presented abroad.

Journal 1: International Journal on Recent and Innovation Trends in Computing and Communication, 11(7s), 01–07, E-ISSN:2321-8169 (Scopus, SJR 0,109)

Citing author: Selvi, S. E.

Title of the article: Geo-Distance Based 2-Replica Maintaining Algorithm for Ensuring the Reliability Forever Even During the Natural Disaster on Cloud Storage System

Link: https://doi.org/10.17762/ijritcc.v11i7s.6971

Impact factor: 0.71

Link for the impact factor: https://www.resurchify.com/impact/details/21101089961

Journal 2: International Journal of Intelligent Systems and Applications in Engineering, 11(3), 769–774, ISSN: 2147-6799 (Scopus, SJR 0,209)

Citing author: Selvi, S. A. E.

Title of the article: A Neoteric Geo-Distance Based 2-Replica Placing Algorithms on Cloud Storage System

Link: https://ijisae.org/index.php/IJISAE/article/view/3283

Impact factor: 1.03

Link for the impact factor: https://www.resurchify.com/impact/details/21101021990

IMPLEMENTATIONS OF THE PROPOSED METHODS

The hybrid model for structuring, storing, and processing distributed data on the Internet has been implemented in an experimental web-based data and information management system. It allows users to organize and manage data according to their preferences, even without any prior experience in database management. The main features of the system include:

- **Flexible table and attribute management** – offers a high level of personalization and control over data;
- **Creation of an unlimited number of tables** – the system allows configuring visibility, defining attributes, and entering data into the tables, with access controlled based on user permissions;
- **Management of individual and collaborative workflows** – suitable for a variety of applications, making it versatile and beneficial for a wide range of users;
- **Easy adaptation to different scientific and work processes** – does not require modifications to the basic structure or functionality when integrated into new projects;
- **The system is designed to be highly customizable**, versatile, and adaptable to different workflows, effectively meeting the needs of users.

TITLE: MODELLING OF DATA STRUCTURING, STORAGE, AND PROCESSING OPERATIONS ON THE INTERNET

Author: M. Eng. Velin Sabinov Hadzhiev

ABSTRACT:

The paper examines methods for modelling data structuring, storage, and processing operations, which form the foundation for developing a model that implements a hybrid approach to data management on the Internet.

The proposed model combines various data warehouse architectures, tailored to meet specific technical and business requirements. The data architecture and flow organization are depicted through IDFO diagrams, which illustrate the step-by-step centralization of data structuring, storage, and processing.

The outcomes of the development process include a classification of methods and models, the creation of a generalized model, and activities related to its preparation and analysis to identify its strengths and weaknesses, along with measures to address them.

An experimental web-based system has been developed to implement this model. The system enables users to organize and manage data based on their specific needs and preferences, even without prior experience in database development.

Key words: Data Processing, Data Structure, Data Storage, Data Integration, Data Model, Hybrid Model, Information System, Data Architecture, Data Centre, Data Warehouse, Relational Database, Data Marts, Data Cube, Big Data, Bottom-up Approach, Top-Down Approach, Central Repository, Cloud Computing, Cloud Database.