UDC 681.3.016

Designing Multidimensional Information Systems Using the Data Vault Methodology

Anastasiya V. Demidova[*], **Yevgeny A. Kuznetsov**[†], **Maxim B. Fomin**[*]

[*] *Department of Information Technology*
Peoples' Friendship University of Russia (RUDN University)
6 Miklukho-Maklaya str., Moscow, 117198, Russian Federation
[†] *Department of digital solutions*
Laboratory of New Information Technologies (LANIT)
14 Murmanskiy proezd, Moscow, 129075, Russian Federation

Email: demidova_av@rudn.university, kuznetsovea@lanit.ru, fomin_mb@rudn.university

The method for designing information systems using the "Data vault" modeling technique, which was formalized by Dan Linstedt, is considered. In case of using "Data vault" the information system is based on the classical formulated by Bill Inmon 3-tier architecture approach to data warehouse design. It includes Operational warehouse of data, Data warehouse, and Data marts. This approach makes it possible to build an information system data warehouse with a metadata repository based on the multidimensional principle. The metadata repository is responsible for collecting data, storing data, and presenting data for analysis. The proposed method of describing metadata provides the ability to specify how to calculate the performance indicators used in the data analysis. The "Data vault" approach allows you to design the data warehouse of an information system using a meta-model that is semantically related to the subject domain of the system and is easily rebuilt in the event of changes in the business model of the subject domain. This approach provides an easy way to generate data marts based on OLAP principles. The key moment in the structure of the information system is the way of transition from the "Data vault" model to the multidimensional model of data representation on the basis of associative rules of the relationship between information objects.

Key words and phrases: data warehouse, multidimensional data model, data mart, OLAP, data vault.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the VIII Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Moscow, Russia, 20-Apr-2018, published at http://ceur-ws.org

1. Introduction

The appearance of low-cost high-performance computing systems has made them available to medium-sized enterprises, whose operation is associated with the implementation of a large volume of operations of various types. Such enterprises have a need for low-cost and easy-to-operate information systems that provide the implementation of the tasks of analysis of the activities of enterprises. Such information systems should meet the following requirements:

- the system must process data arising in the course of the enterprise's activities;
- the system should be able to describe and calculate the key performance indicators that are used in the decision-making process for enterprise management;
- the data warehouse metadata structure must correspond to the business processes of the enterprise;
- there should be an opportunity of operative changes in the system with changes in the activities of the enterprise or in the case of changes in the methodology of the analysis of activities.

2. Information system architecture

During the activity of the enterprise heterogeneous information data sets are generated, which are stored in information subsystems that are external to the analytical information system. The task of the analytical information system is to collect these data from external subsystems and to calculate the key performance indicators that can be used in the process of analyzing the activities of the enterprise and in the process of making decisions on the management of the enterprise.

To provide these functions, the information system must contain the following set of subsystems: Data acquisition subsystem, Data storage subsystem, Data representation subsystem, and Subsystem of control [8, 9]. Thus, data storage is separated from the business users and used by them in solving the problems of analyzing data slices. This separation reduces the cost of modification at the business level. At the same time, this approach enables business users to directly manage and modify the virtual layer (self-service BI).

The architecture of the information system that automates the information processes in accordance with the data model described in the metadata repository [11–14] is shown in Figure 1.

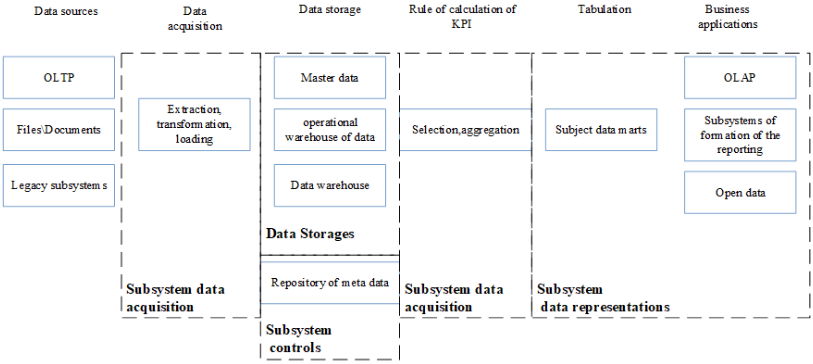

Figure 1. Data warehouse meta-model structure

The analytical information system interacts with external information systems, which are data sources. These are OLTP systems, legacy subsystems, standard format data files, and any other sources of structured data. On the basis of data taken from external sources, the Data acquisition subsystem of the information system forms the correct content of the Operational warehouse of data (OWD). OWD is a storage area in which information exists before it is overloaded in the Data warehouse (DW). DW will combine information related to all aspects of the enterprise. Loading information from OWD to DW is done by normalizing the data according to the rules of the current DW data model.

The calculation of performance indicators is based on data taken from the data warehouse. Performance indicators are placed in special data storage structures — thematic data marts in the Data presentation subsystem. Thematic data mart is a narrow slice of information for users working in one specific task. As a rule, the task of a thematic data mart is to represent data access for business applications [4–6]. For business applications, this means decision support systems that use data representation in the form of OLAP, or subsystems that use a different form of data representation that is convenient for generating reports.

The central block in the structure of the information system is the metadata repository. It is responsible for managing the data model at the meta–model level and is used to manage the process of data movement in the information system. The main requirement for the meta–model is as follows: metadata should be described in such a way that it is possible to specify on its basis the method of calculating performance indicators used in the process of analyzing the activities of the enterprise and in the process of making decisions on the management of the enterprise [7, 10]. From the point of view of business analysts, the most appropriate approach for describing the metadata repository is the multidimensional principle of data organization (metadata as it is data in the metadata repository). Since a multidimensional data model provides a denormalized way of storing data, a "Data vault" model can provide a convenient way of structuring information for the data warehouse. Using the methodology "Data vault" allows you to describe the semantic links of the data warehouse with a description of the information domain of the information system. This provides an opportunity to rebuild the structure of DW in the event of changes in the business model of the subject domain.

3. Description of the data warehouse model using the "Data vault" methodology

One of the ways to build a data warehouse is the data vault methodology. Its use makes it possible to dynamically expand the DW data model without having a complicated task of modifying other subsystems of the information system. The data model must be managed at the meta-model level [15]. The main objects of the meta-model are: a business key (in the terminology "Data vault" — "hub"), a business key transaction (in the terminology "Data vault" — "link") and business key history (in the terminology of "Data vault" — "sat"). Business key is a property of an object that uniquely identifies it within the subject domain. A business key history is a history of changes to object properties that are functionally dependent on that business key. The relevance of the attributes of the dimension is maintained using business key. Business key transaction is a description of the event that occurred between objects that are identified using these business keys [1–3].

As an example of using "Data vault" you can consider the process of on-line sales. The conceptual model of the process is presented in Figure 2.

Figure 3 shows the structure of the DW meta-model as a diagram in the E/R+Merise notation. In order to ensure that the information system modification process does not lose information about the associative links available in the meta–model by link type (aggregation, composition or recursion) and by arity (1:1, 1:N, M:N), this information should be kept by transactions.

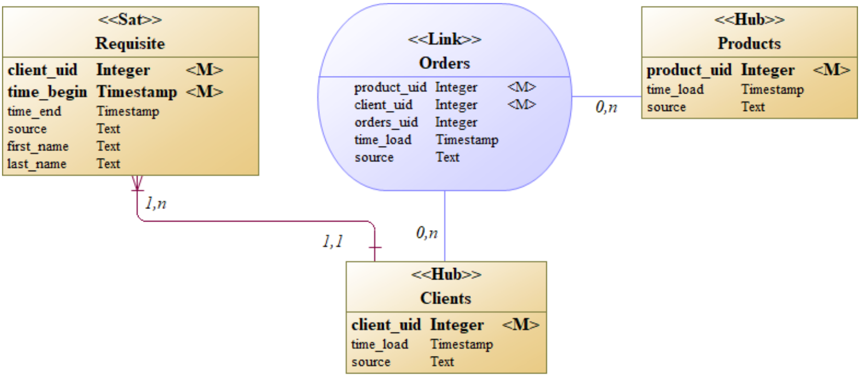

Figure 2. The sales process conceptual model

"Client" act as business keys. "Orders" are transactions between "Clients" and "Products" that implement an association of type "M:N". "Requisite" form the history of the business keys of the "Clients".

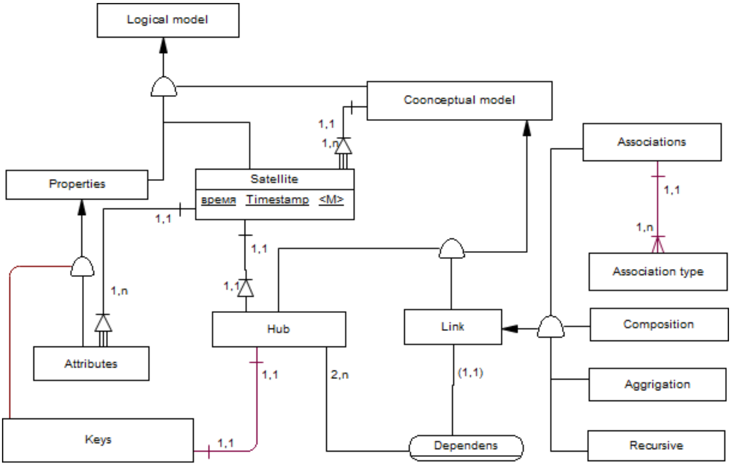

Figure 3. Data warehouse meta-model structure

The metadata repository model is based on the multidimensional data model. This approach makes it easier to establish a correspondence between the metadata and business process parameters of an enterprise, and describes how to calculate the performance indicators and data that are used in the process of completing data marts [16,17]. For the implementation of requests for data must be defined rules of connections (associations) between objects of the multidimensional data model and objects in a “Data vault”. Such rules can be formulated on the basis of the following statements:

1. Within the multidimensional data model in the analytical subsystem, the key can act as a slowly changing dimension;
2. Business key transaction history makes it possible to calculate the values of measures in multidimensional data models used in data marts.

These rules use connectivity at the conceptual and logical levels of representation of the metadata repository model. A complete diagram of the metadata repository model is shown in Figure 4.

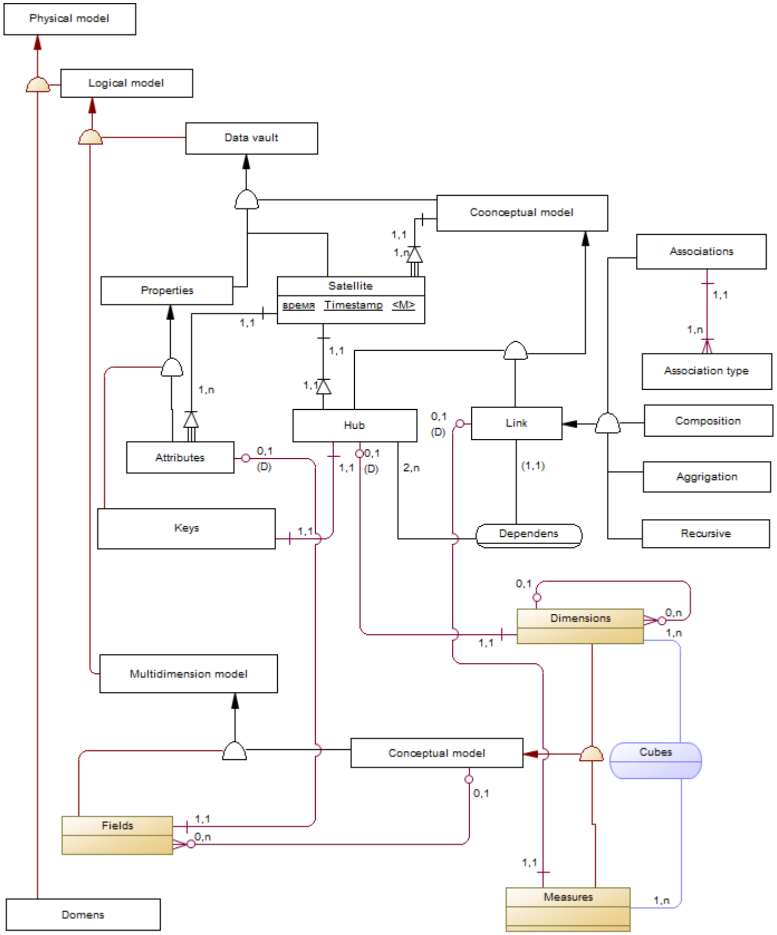

Figure 4. Metadata repository model

4. Multidimensional data model

The structure of multidimensional data model should reflect the aspects of subject domain which are used in the data analysis process. Each aspect corresponds to one dimension of a multidimensional cube H. A full set of dimensions forms a set $D(H) = \left\{D^1, D^2, \ldots, D^n\right\}$, there D^i is i–dimension, and $n = dim(H)$ — dimensionality of multidimensional cube [18]. Each dimension is characterized by a set of members $D^i = \left\{d^i_1, d^i_2, \ldots, d_{k_i})^i\right\}$, there i is a number of dimension, k_i — the quantity of members. Members of D^i are drawn from a set of positions of the basic classifier which corresponds to an aspect of the observed phenomenon associated with D^i [19, 20].

The multidimensional data cube is a structured set of cells. Each cell c is defined by a combination of members $c = (d^1_{i_1}, d^2_{i_2}, \ldots, d^n_{i_n})$. The combination includes one member for each of the dimensions. If the analysis of the observed phenomenon is performed using a large set of diverse aspects, not all member combinations define the possible cells of multidimensional cube, i.e. the cells corresponding to a certain fact. This effect occurs due to semantic inconsistencies of some members from different dimensions to each other and generates a sparseness in the cube.

The complex structure of the compatibility of members may lead to a situation where a certain dimension becomes semantically uncertain if combined with a set of members from other dimensions. In this situation, while describing the possible cell of multidimensional cube the special value “Not in use” can be used to set the member of semantically unspecified dimension.

The subject domain is characterized by the measure values defined in possible cells of the multidimensional cube. The full set of measures composes the set $V(H) = \{v_1, v_2, \ldots, v_p\}$, where v_j is j-measure, p — the quantity of measures in the hypercube. Not all the measures from the $V(H)$ can be defined in the possible cell. This situation can appear in case of semantic inconsistency between the members defining the cell and some measures. While describing multidimensional data cube structure for every possible sell it is necessary to define its own set $V(c) = \{v_1, v, \ldots, v_{p_c}\}$, which consists of certain measures for this cell, $1 \leqslant p_c \leqslant p$. We can use the special value “Not in use” for the description of c measures, which are not included in the set $V(c)$.

5. Conclusions

The paper discussed the method of designing information systems using the methodology of “Data vault”. This approach allows building a data warehouse system based on meta–model, which is semantically related to the subject domain of the system, easily rebuilt in case of changes in the business model of the subject domain, allows you to form multidimensional data marts and calculate the performance indicators of the enterprise.

Acknowledgments

The work is partially supported by the Ministry of Education and Science of the Russian Federation (the Agreement number 02.a03.21.0008).

References

1. D. Linstedt, M. Olschimke, Building a Scalable Data Warehouse with Data Vault 2.0, Elsevier Inc., 2016.
2. W. Inmon, D. Linstedt, Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault, Elsevier Inc., 2015.
3. H. Hultgren, Modeling the Agile Data Warehouse with Data Vault, Brighton Hamilton, 2012.
4. L. Corr, J. Stagnitto, Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema, DecisionOne Press, 2011.

5. W. H. Inmon, Building the Data Warehouse, Wiley Publishing, 2005.
6. W. H. Inmon, Building the Operational Data Store, Wiley Publishing, 1999.
7. W. H. Inmon, D. Strauss, G. Neushloss, DW 2.0: Architecture for the next generation of data warehousing, Elsevier Inc., 2010.
8. R. Kimball, J. Caserta, The Data Warehouse ETL toolkit. Willey Publishing, 2004.
9. R. Kimbal, L. Reeves, M. Ross, W. Thornthwaite, The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. Wiley Publishing, 1998.
10. R. Kimbal, L. Reeves, R. Merz, The Data Warehouse Toolkit: The Complete Guide to Dimensional modelling. Wiley Publishing, 2002.
11. C. Batini, S. Ceri, S. B. Navathe, Conceptual Database Design: An Enity-relationship Approach. Benjamin/Cummings, 1992.
12. S. Singh, S. Malhotra, Data Warehouse and its Methods, Journal of Global Research in Computer Science **2** (5) (2011) 113–115.
13. A. Datta, H. Thomas, A conceptual model and Algebra for On-Line Analytical Processing in Decision Support Databases. Information Systems Research **12** (1) (2001) 83–102. doi:10.1287/isre.12.1.83.9715.
14. C. Fahrner, G. Vossen, A survey of database transformations based on the entity-relationship model. Data & Knowledge Engineering **15** (3) (1995) 213–250. doi:10.1016/0169-023X(95)00006-E.
15. E. Medina, J. Trujillo, Standard for Representing Multidimensional Properties: The Common Warehouse Metamodel (CWM), in: Advances in Databases and Information Systems (ADBIS), Lecture Notes in Computer Science **2435** (2002).
16. V. Jovanovic, D. Jaksic, S. Mrdalj, Data modeling styles in data warehousing, in: Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, Proceedings 6859796, 1458–1463. doi:10.1109/MIPRO.2014.6859796.
17. D. Dymek, W Komnata, P. Szwed, Proposal of a new data warehouse architecture reference model, in: Beyond Databases, Architectures and Structures (BDAS), Communications in Computer and Information Science **521** (2015) 222–232.
18. M. B. Fomin, Cluster method of description of information system data model based on multidimensional approach, in: Distributed Computer and Communication Networks (DCCN), Communications in Computer and Information Science **678** (2016) 657–668.
19. E. Thomsen, OLAP Solution: Building Multidimensional Information System, Willey Publishing, 2002.
20. L. Fu, Efficient evaluation of sparse data cubes. in: Advances in Web-Age Information Management (WAIM), Lecture Notes in Computer Science **3129** (2004) 336–345.