COMPUTATIONAL PROBLEMS OF ELECTRICAL ENGINEERING

Vol. 1, No. 1, 2011

# MODELING OF ENERGY SYSTEM DATASPACE

**Natalia Shakhovska, Mykola Medykovskyi**
Lviv Polytechnic National University,
natalya233@gmail.com

**Abstract -** In this paper the model of an energy system dataspace is described.

**Keywords:** Dataspace, Datawarehouse, Integration Methods, Consolidation, Information Product...

## 1. Introduction

Energy System is a set of power plants, electrical and heating systems and other energy facilities that share a common mode of production, transmission and distribution of electric and thermal energy for centralized management of this regime. However, since the system is dynamic and its elements have evolved at different rates, it complicates the collection and processing of information on elements of such a system. For example, the power stations use variety of software; data from some of the sensors arrive with delay; searching for information in grouped data with power stations and accounting is relevant. To work with different types of information from different sources, we can apply a so-called dataspace. One element of the dataspace is an information product.

Information Product (IP) is a documented information resource, prepared according to the needs of users and submitted as product. Information product can be software, text files, web pages, spreadsheets, xml-files, databases, datawarehouses, etc.

Catalogue of Information Product – metadata about information products – describes the location of an information product, its structure, methods of access to the information resource, etc.

Traditionally, experts used usual for them sources of information for solving tasks [1, 2, 8, 9, 14]. Apparently, this approach has incomplete information, which is processed. Many sources of data and services that exist on the Internet are causing the need for a radical change in the methods of getting data. This change is a task which is formulated independently of existing data sources. After its formulation the identification of relevant sources, bringing data to the appropriate type, integration, identification services which allow solving a separate part of tasks should be carried out.

The adoption of adequate solution require the data, coming from different sources to satisfy the following requirements: be complete, consistent and received on time; be informative, because they should be applied for decision support; be of uniform structure for the opportunity of being downloaded in a single datawarehouse and analyzed; kept in uniform models of data and be independent of the development platform for the opportunity of using this data in other means. But today there are no data processing methods that would satisfy all of the requirements for data processing [6, 7, 12].

The article describes a dataspace, built for an energy system.

## 2. Algebraic system of energy system dataspasce

Datawarehouse is an aggregated information resource that has consolidated information from all areas of concern and is used for decision support.

Dataspace is a set of all information product domains

DS= <DB, DW, Wb, Nd, Gr>,

where DB, DW, Wb, Nd, Gr are information products that submit a set of databases, datawarehouses, web pages, text files, spreadsheets, image data respectively. In an energy system databases, datawarehouses, text files, and spreadsheets that are described in different formats are used.

Talking about an information product, we mean its content (information resource, IR), and the set of information about it (accommodation, access scheme, speed of information update, etc.). We are also interested in the operations which are carried out over IR depending on its DSIR. The main task of a dataspace is to allow a user to work with data sources without knowing its DSIR, accommodation, access methods, etc.

Consolidated data of an energy system is derived from multiple sources and systematically integrated heterogeneous information resources, which together possess such features as completeness, integrity, consistency, and adequacy. This consolidated information is model of the subject area for its analysis and processing efficiency in the processes of decision making.

Information products describe the specific subject area, and consolidated data constitute the dataspace. One of the problems that interfere in the process of consolidation, is the uncertainty of data, the result of duplication, inaccuracies, data absence, contradictions of the data (Fig. 1).

Another problem is the determination and approval of the schemes of the data of information resources. The existing methods, as shown in the section, are working either on the known data schemes, the data source (information products) are under strict control, which makes it impossible to set various semantic relationships [7, 8, 9, 11]. Also one of the obstacles which is analyzed using the integration is that developers have not always adhered to the standards in the development schemes of data. The analysis of capabilities of the existing standards has shown that the development of a data dictionary allows one to avoid this problem and to unify partially the scheme of data sources.

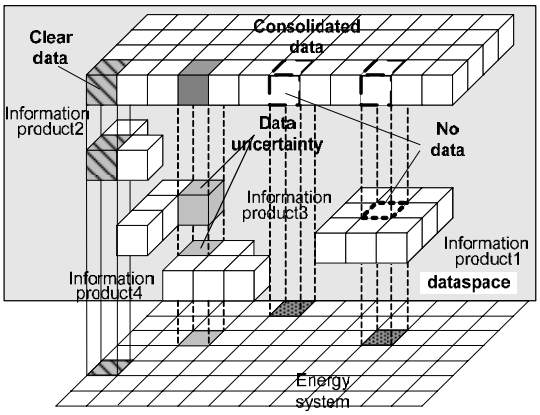


*Fig. 1. Consolidated data and dataspasce*

To implement storage and dataspaces using database management systems, means of data exchange and integration are necessary. Data sources such as spreadsheets, multimedia information, etc, that are used in energy systems, can have their own means of storing and processing, and then the task of integration is the recognition of these information resources and access to them. When talking about data storage, the structure of sources is known in advance, and the main challenge is clearing and loading data itself. For spaces of data it is necessary to provide the opportunity to work with the software, which theoretically might not be at the user's workstation. If such a possibility has been not foreseen, it is necessary to predict the development of such data storage structure so that it can retrieve data from data sources to provide answers to user queries.

Information product state $S_{Ip}$ comprises its information resource *IR* fixed at the given moment and the corresponding information about the information product (data catalogue) *Cg*:: $S_{Ip} =< Ir, Cg >$.

**Dataspace state** ($S_{DS}$) is the set of states of all information products of subject area and relations between them.

The set of dataspace information products, operations over IR in them and predicates on the set of IR are called the **dataspace class algebraic system**.

$$\mathbf{DS=<Ip, \Omega_P, \Omega_F>,} \quad (1)$$

where **Ip** = *DS* is the set of information products of the subject area (database **DB**, Data Warehouse **DW**, static Web-pages **Wb**, text data **Nd**, graphics and multi-media data **Gr**),

$\Omega_P = \{O_{P0}, \boldsymbol{O_{Pu}}, \boldsymbol{O_{Pb}}\}$ – the set of operations on information resources IR, where: $O_{P0}$ – null-operation, which results in a given state of IR in the data space; $\boldsymbol{O_{Pu}}$ – the set of unary operations on data space DS. The result of these operations is the change of the data space state; $\boldsymbol{O_{Pb}}$ – the set of binary operations on the data space. The result of these operations is the formation of the new data space.

$\Omega_F$ – the set of predicates defined on the set of information products of the dataspace.

The result of a nular operation on dataspace is the state of the information product *Ip*: $S_{Ip} = O_{P0}(DS, Ip)$.

Unary operations over the data spaces are:

$$\boldsymbol{O_{Pu}} = \{Cons, Se_{sim}, Se_{st}, Se_m, \sigma_{ac}, Agent, Ag\}$$

where $Agent$ – Operation of IRDS definition;

$Se_{sim}, Se_{st}, Se_m$ – Search Operations;

$\sigma_{ac}$ – Access Operation.

IRDS determination is carried out by using intelligent agent and means the addition into the *Cg* the new data about IP IRDS: $f_{Ip}(DS) \xrightarrow{Agent} Cg \bigcup Ip.Cg$, where *Cg* is data space catalogue, *Ip.Cg* – IP catalogue Ip.

The agent is

$$Agent = < \boldsymbol{Cg}, EM, Dic, ExB, Sol, Eff >,$$

where ***Cg*** is information about sources that are already in the *DS*;

*EM* – a component of the agent responsible for the perception of the environment, which is the environment of model management;

*Dic* – the synonymic terms that indicate the sources of the same properties;

*ExB* – the base of agent experience containing "the history of impacts" on the agent from the environment and the corresponding agent reaction;

*Sol* – the component that is responsible for training;

*Eff* – the component responsible for the actions of the agent.

*Data Integration* is the association of IP information resources in the local datawarehouse of defined structure DW.rel for further processing for decision-making management:

$$DW.rel=<Ip_1.Ir \cup \ldots \cup Ip_n.Ir;$$
$$Ip_1.Cg \cup \ldots \cup Ip_n.Cg> \xrightarrow{cons} S_{DS}.$$

*Data Aggregation* is the calculation of generalized values based on the dimensions of relationships to support strategic and tactical management with detailed data: *rel = Ag* (DB1.***r***, ..., DBn.***r***).

*Arbitrary Data Request* – users must be able to query any data element, regardless of its format and data model. It is carried out on the keyword and keyword IP Cg catalogue: $Se_{sim} : \sigma_{key}(Cg)$.

*Structured Queries* - are built using SQL and similar languages. With the help of catalogue it is determined whether the source in which the search is carried out contains structured information. The query is conducted directly to the data source.

$$Se_{st} : \sigma_{Cg.x='st'}(\pi_x(\sigma_{key}(Ip_1)) \cup ... \cup \pi_x(\sigma_{key}(Ip_n)))$$

Requests to metadata should be provided with opportunities of: obtaining data about the source of answers and the source location; identification of data elements in the dataspace that can vary by a given data element and hypothetical queries support; determination of the level of uncertain response $\mathrm{Se}_{\mathrm{m}} : \sigma_{use}(Cg)$, where use is the set of user preferences (query requirements), its profile, or demands, which relate to the decision.

Dataspaces of an energy system can be put one into another.

Binary operations on IP sets are advanced set-theoretic union and intersection operations.

Binary operation of data spaces *union*:

$$DS_3=DS_1 \cup DS_2: \text{profile}(Agent(Cg_1) \cup Agent(Cg_2)),$$
$$Cg_3 = Cg_1 \cup Cg_2.$$

Binary operation of the data spaces *intersection*:

$$DS3=DS1 \cap DS2:$$
$$\text{profile}(Agent(Cg_1) \cap Agent(Cg_2)),\ Cg_3 = Cg_1 \cap Cg_2.$$

Advanced operations of union and intersection are the set-theoretic unions or intersections of data space catalogues. This user's access to IP from data spaces with $DS_1$ and $DS_2$ is determined by the profile formed on the basis of a new catalogue $Cg_3$.

Predicate on information products is IP Registry, which contains the most basic information about each of them: the source, name, location in the source, size, creation date and owner, etc., as well as the results of comparison of the similarity of data structures with each other. In order to distinguish sources glossary of terms and concepts (keywords) *Dic* is used, that contains the synonymous description of the same concept in the different data sources.

Data dictionary filling is conducted at the beginning by using the developed ontology domain, then – automatically: Metadata*(DS)* $\cup$ *Dic* $\Rightarrow$ *ODW*.

*Null-predicate* $\Omega_{F0}$: returns TRUE, if for the given information product *Ip* the IP data structures are known, and FALSE otherwise.

*Comparison Predicate* of the information Resources IP Data Structures: $\Omega_{eq}(Ip_1, Ip_2) \rightarrow Dic$.

### 3. Energy system dataspasce implementation

The power plants, electrical and heating systems use databases for information saving. Comparative characteristics of datawarehouse implementation are given in Table 1.

To integrate the geo-databases and relation database specialized tools that convert vector data into a special format are used. The dataspace requires much more technological and methodological solutions, as it contains processed information from various data structures, uncertain advance, and use different tools for processing and storage.

If we look at technologies that will help realize the potential of data space, we first need to focus on grid and cloud computing.

Service-oriented grid-technology provides new opportunities that were not in the networks of organized scheme peer-to-peer or client-server.

The functioning and interaction of services similar to the technology of multi-agent systems, intelligent agents perform the role of grid services. Thus, in this view, they have several advantages over web-services. Among them: the possibility of data retrieval functionality is not confined to a set of procedures implemented on the server data storage, the ability to analyze both globally and in corporate networks, to continue service agents search and data collection even after a specific request; built-in possibility of transfer of data access from user to the entire sequence of grid-services using digital certificates.

Service-oriented approach to weakly structured data as a data space can already create a new level of services that operate not only databases or metadata, but also work directly with Web-data and other weakly structured resources. At the same time you need the application of revised database technology for data space that allows the amount of new positions to solve the problem of

heterogeneous integration of federal information resources.

**Cloud computing** is the technology of data processing, where software is provided as a user of Internet services. The user has access to data, but cannot control and should not care about the infrastructure, operating system and proprietary software, with which it works.

*Table 1*

**Comparison of datawarehousing implementation**

| Tool | Advantages | Failings |
|---|---|---|
| Oracle Warehouse Builder, Oracle Data Integration, Oracle Optimized Warehouse, Hyperion | Corporations level DBMS; it can be used as a component oriented to the data of architecture in the SOA or BI environment. Includes: data movement, synchronization, verify-cation of data quality, data management | For integration first describe the data source and manually set up procedures to check data quality |
| Database Application Server | A platform for creating and deploying multi-player network programs for the Web, where customers can be both standard browsers, and Java-applications | Middleware software |
| SQL Server 2008 | Contains Integration Services, Analysis Services, Reporting Services, Management Studio, Business Intelli-gence Development Studio | As for Oracle, works with pre-known sources |
| Biz Talk | Integration server allows analyzing text data and recording to the data store, operates at both indoor and partially interoperated level | Operates on the principle of no-tifycation, so the quality de-pends on the integra-tion of user |
| Netezza | Massively parallel without resources divisi-on architecture (storage le-vel) and symmetric multi-processing architecture (host level), acelerating of data processsing while transferring to storage | Closed technology, hundreds of terabytes scalability, insufficiently developed integration tools |
| Teradata | Due to the mass paralle-lism technology the Tera-data storage can be scaled to multiple petabytes | Closed architecture, does not support SaaS |

Several options may form the provision for the use of computing power and databases datacenter: Saa, Paa, Haa, Iaa, Caa - as Internet services. S, P, H, I, C – Software, Platform, Hardware, Infrastructure, Communication – respectively, the software, platform, hardware, infrastructure, communications. Platform paradigm of Cloud Computing consists of components: virtualization, SAAS, SOA (Service-Oriented Architecture) and SS (Software Services).

Today, there are such types of cloud computing:

a) Windows Azure is a cloudy operating system produced by Microsoft, designed for developing and running web-applications that run on the server provider, not on your computer. It also includes the platform of Microsoft Azure. It uses relational structures for data storage.

b) Google App Engine – a platform that allows one to use infrastructure for creating and hosting their applications. It uses non-relational distributed data warehouse.

Let us compare both technologies. In the cloud, as in the grid, powers are accumulating to get more economical, extensive decision and deprive the user of having to work exclusively on their own hardware and software. A business model for a typical grid can be called design and project-oriented. Companies transfer their resources under the administrative control grid, creating a pool of distributed resources, the physical organization of which can be generally unknown to the user. The user data he/she needs receives in the form of service, perhaps even more than the amount of resources, which he/she can calculate.

Grid Architecture and the clouds are different because they were created according to different conditions. The former was by distributed computing resources influenced the desire to use expensive more efficiently, making them dynamic and homogeneous. So the architecture is built on the integration of existing resources, including hardware and software, operating systems, local tools that provide management and security. The result creates a "virtual organization", resources which, translated into a logical form, may be used only by members of the organization.

Cloud architecture is opened for access through the network, not only within the grid. Reference to pools with computing resources and storage data systems are made by the standard protocols, such as WSDL and SOAP, or using technology Web 2.0 (REST, RSS, AJAX), as well as through the existing technology grid.

### 4. Analysis of accumulated complete descriptions of objects

For the dataspace of an energy system the data completeness and usefulness of objects of accumulated decisions are affecting the quality of the dataspace. Completeness collected descriptions of objects – the relative count of objects or documents available in the

data source to the total count of objects, which hit the local store (illustrated in Fig. 2).

In the submitted chart analysis of the algorithm is modified consolidation. This algorithm is compared with the work of not modified integration algorithm which it use in the Oracle Data Integrator. Data come from the local repository database of travel agencies. Data structures of those sources are unknown. Count of input records databases that get in the local repository of data is 15000. Without a prior determination of the structure of the data sources was not loaded in the repository of consolidated data (count of absent data in the chart), some data could not be loaded because of the discrepancy between the structure of local storage and structure of sources (uncertain data in the chart). Unmodified consolidation also does not include data that have arrived too late (data inserted during the operation of integration). Apparently, the number of tracks using the modified consolidation is the largest.

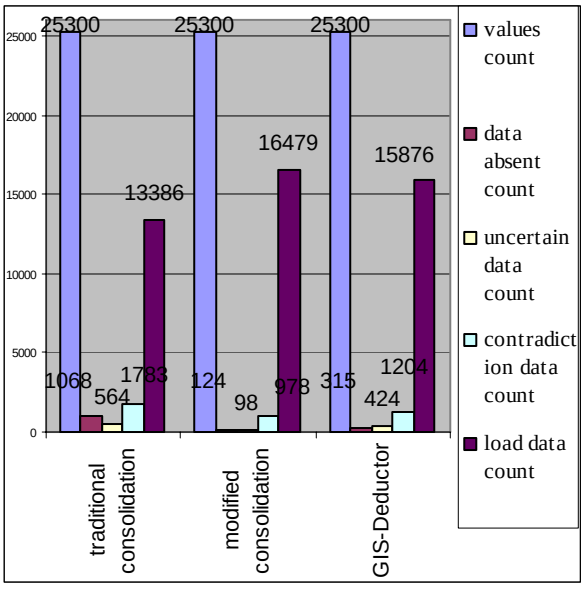


*Fig. 2. Analysis of accumulated complete descriptions of objects.*

## 5. Conclusion

In this article the formal model of data space is presented. It is shown that the algebraic system of database and data warehouse class is the subclass of the data space class algebraic system. The dataspace of an energy sector is described. **Scientific novelty.** The novelty lies in the introduction of data space class algebraic system. **The practical value** lies in the identification of the main tasks and components of the data space and relations between them. **Further investigation** will cover the formalization of search methods of the unstructured, semi-structured and strictly structured data and building the appropriate algorithms.

## References

1. Turban, E. Decision support and expert systems: management support systems. -Englewood Cliffs, N.J.: Prentice Hall, 1995.

2. К. Дрюэк. "Хранилища данных: сходство и различия подходов Билла Инмона и Ральфа Кимболла", 2005, http://www.b-eye-network.com/view/743

3. Dan Linstedt. Data Vaulttm overview the next evolution in data modeling. – 2005, http://www.tdan.com/i021hy01.htm.

4. Огляд технологій інтеграції інформаційних систем, 2006, http://www.microsoft.com/Ukraine /Government/ Analytics/ IntegrationTechnologies/Overview.msp

5. С. Кузнецов. Пространства данных: исследова тельский полигон или путь к новому поколению систем управления данными? http://synthesis.ipi.ac.ru/sigmod/ seminar/s20060420

6. D. Kossmann, J.-P. Dittrich. Personal Data Spaces. http://www.inf.ethz.ch/news/focus/res_focus/feb_2006/index_ DE.

7. Рогушина Ю.В., Гладун А.Я. Формирование тезауруса предметной области как средства моделирования информационных потребностей пользователя при поиске в Интернете //Вестник компьютерных и информационных технологий. – М.: 2007. – № 1. – С.26–33.

8. ETH - Databases and Information Systems – iMeMex, www.dbis.ethz.ch/research/current_projects/iMeMex

9. Processing of natural language queries to a relational database. Samsonova M, Pisarev A, Blagov M, http://www.cs.dartmouth.edu/~brd/Teaching/AI/Lectures/Sum maries/natlang.html

10. Основные концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных. http://www.citforum.ru/ database/articles/search_sys.shtml

11. Особенности построения хранилищ данных. http://citforum.uar.net/seminars/cis99/sch.shtml/

12. Kacprzyk J., Ziolkowski A. Database Queries with Fuzzy Linguistic Quantifiers // IEEE Transactions on Systems, Man, and Cybernetics. SMC-16, 1996. – P. 512-529.

13. Fuzzy Grouping в Microsoft SQL Server 2005 http://msdn.microsoft.com/msdnmag/issues/05/09/SQLServer2 005/default.aspx.

14. Тенденции в области Хранилищ данных на 2007 год / [Електронний ресурс] / TWAN. – 2007. — Режим доступу: http://citcity.ru/15272/.

15. Б.И. Плоткин «Универсальная алгебра, алгебраическая логка и базы даннях»: - М.:Наука, 1991. – 448 с., ст. 292.

## МОДЕЛЮВАННЯ ПРОСТОРУ ДАНИХ ЕНЕРГЕТИЧНОЇ СФЕРИ

Н.Шаховська, М.Медиковський

Описана модель простору даних енергетичної сфери.

**Natalia Shakhovska** – Ph D., Associate Professor,. Research investigations: datawarehouses, databases, dataspaces, integration systems.

**Mykola Medykovskyi** - DSc, Professor, Research investigations: energetics, mathematical modelling.

Director of Institute of Computer Science and Information Technologies, Dean for Undergraduate Education, Deputy Director