ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 72
IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine" (ITSMSSM 2017)

Review and analysis of means and methods for automatic data extraction from heterogeneous sources

Alexey Samoylov, Alexey Tselykh, Nikolay Sergeev , Margarita Kucherova
Institute of Computer Technology and Information Security
Southern Federal University
Taganrog, Russian Federation
{asamoylov, tselykh, nesergeev, mkucherova}@sfedu.ru

***Abstract* — There is a problem in the process of data analysis, which is related to their extraction and preparation. This problem is the consequence of a necessity for integration of heterogeneous structures both in structure and format. The technical solution to this problem is to use ETL-systems that automate processes of extraction, transformation and loading of data into a storage according to strictly defined rules. To date, scientific research in this area focuses on increasing performance and documenting the semantics of the process for its reuse. The paper presents results of a review and analysis of actual solutions in the field of extraction of heterogeneous data of large volume.**

Keywords— Big data, data extraction, knowledge discovery, data mining, heterogeneous data sources

I. INTRODUCTION

The One of the most promising areas of research today is data analysis. The objectives of the analysis may be of a different nature - social, economic, political, etc. However, in spite of the subject area and the nature of the source data, similar problems can be observed. In modern realities, such common problems have been referred to as the "three V problem" (volume, velocity, variety) [1], which pertains to the Big Data sphere. This problem is extremely urgent, as evidenced by the research estimate - on average, up to 70% of resources (financial and time) are spent on data collection, preparation and downloading from the total amount of work with data [2-4].

Let's give a brief description of each of the problems that make up the three V. The volume problem is associated with the high growth rates of accumulated data, due to the widespread introduction of information technologies and the growth of their accessibility for the population. This problem generates high-speed read / write tasks, machine processing and centralized storage of hundreds of gigabytes of information [5].

Closely related to the problem of volume is the problem of the rate of change (velocity) of data, which is expressed in the speed of growth of the data to be processed. This problem puts even more stringent requirements on algorithms for reading, writing and processing data.

The problem of diversity is the least elaborated, both scientifically and technically. The diversity of data is a consequence of the heterogeneous nature of the sources. For example, the analysis of public reaction to resonance events requires sampling data from various open sources (social networks, blogs). Obviously, such sources will differ in the format and method of presentation of information (tweet, post on the social network, personal blog entry, etc.).

In addition, in individual cases, the individual features of the domain are added to the problem of three Vs. So, the research conducted by the team of authors of the paper is connected with social modeling. This causes, in addition to the above three V, specific problems: identification of "useful" data sources, calculation of key indicators by indirect characteristics. These areas to date are represented preferably by private solutions obtained heuristically and their analysis is not possible.

Having abstracted from particular problems, you can see that the solution to the problem of the three V is inextricably linked to ETL-systems (extract-transform-load). With their help, a data specialist builds the process of integrating heterogeneous sources. Modern research in this area focuses on obtaining particular solutions for individual stages of the ETL process: improving productivity, documenting semantics, and automating design. In this paper, we will consider the main works devoted to the problem of extracting, transforming and loading data from heterogeneous sources.

II. DOCUMENTING THE SEMANTICS OF ETL TASKS

The most interesting from the point of view of carrying out scientific research and promising, according to the authors, is semantic modeling. This approach involves the use of ontological models and semantic WEB principles for describing ETL processes.

At the same time, researchers in this field are divided into two camps: the first uses semantic data modeling to increase the "readability" of the model and simplify its modification, the second - to model the data extraction processes in order to automate the construction of ETL-procedures.

One of the most advanced in the field of semantic data modeling is the Semantic-ETL framework, presented in the works of Bansal et.al. [6,7]. This framework modifies the process of converting heterogeneous data by building a semantic model that establishes relationships between objects in the source data models and objects in the data model of the target storage (Fig 1).

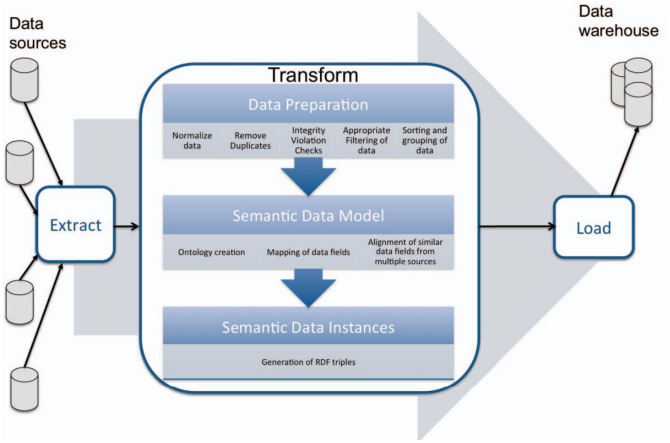

Fig. 1. The conceptual model of Semantic-ETL framework

The main technologies of the framework are RDF and SPARQL, which increases the possibility of distributing this solution in the commercial sphere. It should also be noted that the processes of downloading and extracting, as well as preparing data at the transformation stage, remain classical.

Semantic modeling of data extraction processes is presented in [8,9]. This approach uses technology also uses the ideas of semantic web and ontological modeling (OWL-DL) to describe data extraction processes (fig. 2).

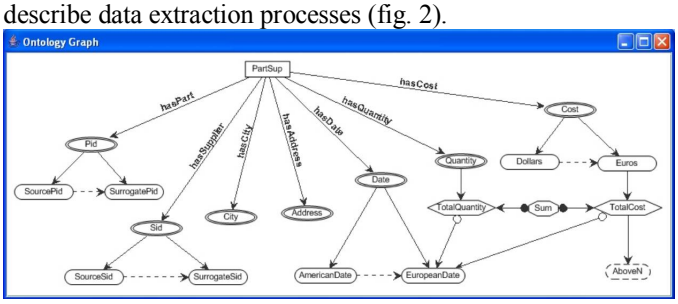

Fig. 2. Ontological modeling of ETL-processes

The main idea of the authors is to semantic annotation of data sources and data stores for subsequent comparison. By searching inter-attribute semantic comparisons, the authors build an ontological model, which then generates an ETL procedure. During the generation process, the descriptive logic and reasoning mechanisms included in the OWL-DL.

A special feature of this approach is the conceptual representation of the generated operations, which still requires the attention of the data specialist.

In the works of A. Azzini et.al. [10] presents an alternative approach to the semantic modeling of ETL-processes, based on the technology of process mining. This technology is used to calculate discrepancies between data from heterogeneous sources. With its help, the authors automate the process of extracting the source data and reduce errors in the conversion and loading process.

In the field of semantic data modeling, approaches based on various private techniques are also known: the algorithm for detecting influence-based modules (IBMM) [11], online extraction of associative rules [12], graph analysis [13], data origin analysis [14]. Common to all considered approaches is the fact that they cover one of the three phases of the ETL process. And at the moment complex solutions based on semantic technologies are absent.

III. GRAPHICAL MODELLING OF ETL TASKS

Like any process related to computer technology, ETL has received many different visual modeling methodologies. It is noteworthy that in this area researchers repeat the development of technologies for automation of design and programming.

One of the first works in the field of modeling the extraction, transformation and loading of data is the paper by P. Vassiliadis et.al. [15], which describes the ARKTOS II framework. In this framework, the author's graphic notation is applied (fig. 3), which allows to describe the processes of filling data stores.

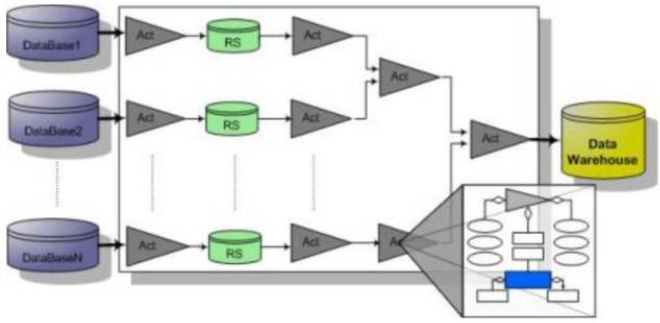

Fig. 3. The graphical notation of ARKTOS II framework

The notation and framework proposed by P. Vassiliadis et.al. at the time of its publication covered all stages - from the description of data sources to the modeling of the storage structure, and also included the processes of data analysis and their transformation. However, to date, the degree of heterogeneity of sources is much higher, which makes it impossible to apply this framework to solve the problem of three V.

In modern conditions, approaches to modeling ETL-tasks were divided into four key concepts

1) object-oriented;
2) process;
3) domain-specific languages (DSL);
4) meta-modeling.

Object-oriented solutions can be found in [16]. These approaches have brought to the world of data analysis all the known advantages and disadvantages of the UML language - the visibility of models, excessive diversity, high degree of abstraction, the complexity of the transition to computer-processed algorithms.

Process approaches are based on using BPMN notation, which, due to the BPEL language, is more productive, but at the same time a less readable way of documenting ETL processes. This approach is found in [17, 18]. A feature of the process approach in data analysis is the consideration of ETL tasks as business processes. This makes it possible to reproduce well-tuned tasks repeatedly and is not applicable in tasks with a dynamically changing set of sources (both in composition and in the format of presentation of source data).

Domain-specific languages, as well as in the field of programming and developing information systems, are of a highly specialized nature and are used in individual classes of ETL-tasks. Therefore, in [19] the add-in over the BPMN language is presented, which allows to build data extraction, transformation and loading processes in terms of data mining (Fig. 4).

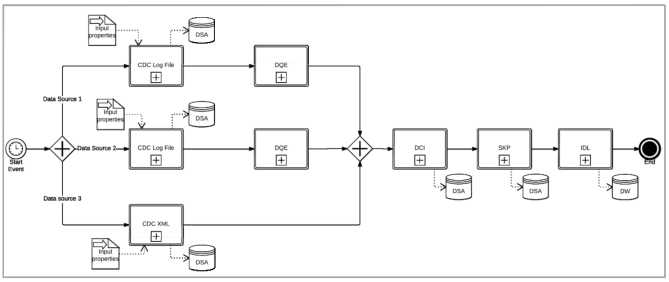

Fig. 4. Domain-specific language for ETL-tasks based on BPMN

Like other works in the field of domain-specific languages, this study is limited to the capabilities of the BPMN-pattern library and does not allow describing all the variety of problems of extracting heterogeneous data.

The most new means of graphical description of software systems is the meta-modeling approach. The basic goal of this approach is to avoid the restriction of object-oriented and domain-specific languages by providing the possibility of expanding the alphabet and rules for constructing models.

In the field of meta-modeling of ETL-procedures, it is worth mentioning the paper [20], in which MDD (Model-driven development) technologies are used. This technology is based on the construction of a number of interrelated models, the transformation between which allows abstracting or refining the technological aspects. In this context, meta-modeling of ETL-procedures is an effective tool for transferring procedures for extracting, converting and loading heterogeneous data onto new technological platforms.

Unlike the documentation of semantics, the approaches to the complex description of ETL problems are presented in the field of graphic modeling, however their scope of application is essentially limited, and the resulting benefits from the application are not high.

IV. AUTOMATED GENERATION OF ETL TASKS

Another approach to constructing procedures for extracting heterogeneous data is automated generation. In this area, the differences between the existing approaches lie in the initial data, which uses the so-called generator. There are the following main directions in this area:

1) Generation on the basis of libraries of standard components;

2) Generation based on machine learning, natural language processing technologies;

3) Generation based on requirements analysis.

Generation on the basis of libraries of typical components is presented in the works of Knap T. Et.al. [21], Schultz A. et al. [22]. Both researchers offer a similar look at the extraction, transformation and loading of data: each process consists of a finite set of DPT (Data processing tasks) tasks, and each task consists of many simple DPU (Data processing unit) operations. In the works of researchers, the corresponding frameworks are proposed, which have replenished libraries of tasks and operations. Simulation is allowed at the task level as well as at the level of operations (if necessary, modification of the basic task).

Generations based on machine learning are devoted to the work of Khouri S. et.al. [23], the main difference from the above-described frames is the automatic formation of a procedure based on the library of finished operations, based on the analysis of data schemes of sources and storages.

Requirements-based generation is devoted to the work of Romero et.al [24], which, through natural language processing methods and parsing of XML documents, formalizes the requirements and, on their basis, performs semiautomatic generation of the ETL procedure.

Regardless of the method, the approaches based on generation have a significant disadvantage - the limited libraries, on the basis of which the assembly of the finished ETL procedure takes place.

V. CONCLUSIONS

The paper examined and analyzed the main means and methods of extracting data from heterogeneous sources, described in the modern literature.

The general conclusion from the analysis is that at the moment there is no comprehensive solution that completely covers all stages of filling the data warehouse with information for analysis. Thus, the most promising semantic approaches are limited to one of the three stages (extraction, transformation, loading) of the process of filling the storehouse with data from heterogeneous sources.

Approaches based on graphic modeling, ultimately represent abstract solutions that require the attention of a data specialist and do not significantly reduce the cost and complexity of the ETL phase.

Approaches based on generation are sensitive to changes in the domain and the laboriousness of filling libraries are comparable to the manual development of the ETL procedure.

Thus, if it is necessary to build a procedure for extracting data from heterogeneous sources, it is necessary to combine existing approaches, which should take into account the degree of obsolescence of technology, the speed of change in data sources and analysis tasks, as well as the qualifications of specialists involved in this task.

ACKNOWLEDGMENT

The study was carried out at the expense of the internal grant of the Southern Federal University No. ВнГр-07 / 2017-28

REFERENCES

[1] Simitisis A. et al. Data warehouse refreshment // Data Warehouses Ol. Concepts, Archit. Solut. 2006, pp. 111–134.

[2] Kimball R. et al. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data // Wiley. 2004, 526 p.

[3] Inmon W.H., Strauss D., Neushloss G. DW 2.0: The Architecture for the Next Generation of Data Warehousing // DW 20 The Architecture for the Next Generation of Data Warehousing. 2008, 400 p.

[4] Victor Mayer-Schoenberger, Kenneth Cuciere. Great data. A revolution that will change how we live, work and think = Big Data. A Revolution That Will Transform How We Live, Work, and Think. With the English. Inna Gideuk. - Moscow: Mann, Ivanov, Ferber, 2014, 240 p. - ISBN 987-5-91657-936-9.

[5] Romero O., Simitsis A., Abelló A. GEM: Requirement-driven generation of ETL and multidimensional conceptual designs // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2011, vol. 6862 LNCS, pp.80-95

[6] Bansal S.K. Towards a Semantic Extract-Transform-Load (ETL) framework for big data integration // Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014. 2014, pp. 522-529

[7] Bansal S.K., Kagemann S. Integrating Big Data: A Semantic Extract-Transform-Load Framework // Computer. 2015, vol. 48, № 3 , pp. 42-50

[8] El Akkaoui Z. et al. BPMN-based conceptual modeling of ETL processes // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2012, vol. 7448 LNCS, pp. 1-14

[9] Kabiri A., Wadjinny F., Chiadmi D. Towards a framework for conceptual modeling of ETL processes // Communications in Computer and Information Science. 2011, vol. 241 CCIS.

[10] Schultz A. et al. LDIF - Linked Data Integration Framework // Proc. 11th Interational Semant. Web Conf. ISWC2011. 2011, Vol. 782, pp. 1–6.

[11] Schultz A. et al. LDIF - A Framework for Large-Scale Linked Data Integration // 21st Int. World Wide Web Conf. (WWW 2012), Dev. Track, Lyon, Fr. 2012, pp. 1–3.

[12] Knap T. et al. UnifiedViews: An ETL framework for sustainable RDF data processing // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014, Vol. 8798., pp. 379-383

[13] Knap T. et al. UnifiedViews: Towards ETL tool for simple yet powerfull RDF data management // CEUR Workshop Proceedings. 2015, vol. 1343, pp. 111-120

[14] Shigarov, A.O., Paramonov, V.V., Belykh, P.V., Bondarev, A.I.. Rule-based canonicalization of arbitrary tables in spreadsheets // Communications in Computer and Information Science. 2016., V. 639, pp. 78–91

[15] Oliveira B., Belo O. A domain-specific language for ETL patterns specification in data warehousing systems // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015, vol. 9273.

[16] Elleuch N., Khalfallah A., Ahmed S. Ben. Software Architecture in Model Driven Architecture // 2007 Int. Symp. Comput. Intell. Intell. Informatics. 2007, pp. 219–223.

[17] Kang K.C., Lee J., Donohoe P. Feature-oriented product line engineering // IEEE Softw. 2002, Vol. 19, № 4, pp. 58–65.

[18] Jifeng H., Li X., Liu Z. Component-Based Software Engineering // Theor. Asp. Comput. 2005. 2005, pp. 70–95.

[19] Legner C., Heutschi R. SOA Adoption in Practice - Findings from early SOA Implementations // ECIS. 2007. № 2007, pp. 1643–1654.

[20] Khouri S., Abdellaoui S., Nader F. Avoiding ontology confusion in ETL processes // Communications in Computer and Information Science. 2015. Vol. 539. , pp. 119-126

[21] Nabli A. et al. Two-ETL phases for data warehouse creation: Design and implementation // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. Vol. 9282. , pp.138-150

[22] Romero O., Simitsis A., Abelló A. GEM: Requirement-driven generation of ETL and multidimensional conceptual designs // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2011,Vol. 6862 LNCS.

[23] Rogozov Yu.M. I. Interdisciplinary paradigm of organization of systems // Proceedings of the VII International scientific and technical conference "Technologies of development of information systems (TRIS-2016)". T.1. - Taganrog: Publishing house SFU, 2016. - P. 3-12.S. Kucherov, Y. Rogozov, A. Sviridov The Model of Subject-Oriented Storage of Concepts Sense for Configurable Information Systems // Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16) Volume 1 in Advances in Intelligent Systems and Computing, , 2016 ,vol. 450, pp 317-327

[24] Kucherov, S., Rogozov, Y., Sviridov, A., and Rasol, M. 2015. Approach to data warehousing in configurable information systems based on action abstraction / Communications in Computer and Information Science, 2015, vol, 535, pp. 139-148

[25] Kucherov S., Rogozov Y., Sviridov A. NoSQL approach to data storing in configurable information systems // Communications in Computer and Information Science, 2016, vol 584, pp.120-134

[26] Kucherov, S., Rogozov, Y., Sviridov, A. / SUBJECT-ORIENTED DATA MODELLING APPROACH// 16th International Multidisciplinary Scientific GeoConference SGEM 2016, www.sgem.org, SGEM2016 Conference Proceedings, ISBN 978-619-7105-58-2 / ISSN 1314-2704, June 28 - July 6, 2016, Book2, vol. 1, pp 437-444

[27] Kucherov, S., Rogozov, Y., Sviridov, A./ The subject-oriented notation for end-user data modelling //10th International Conference on Application of Information and Communication Technologies, AICT 2016 - Proceedings, pp. 358-362