The Integrated Enterprise Data Warehouse Engineering

Sergey Zykov
TEKAMA

Abstract (English)

The paper presents a novel engineering technology for integrated heterogeneous enterprise data warehouses. The enterprises have accumulated a huge and rapidly increasing data bulk, which is difficult to manipulate due to its heterogeneous nature. The integrated software engineering technology unifies manipulation of heterogeneous enterprise data warehouses. The technology incorporates mathematical models, methods and tools based on portal architecture

Key words: *data warehouse, software engineering, enterprise software system*

1. Introduction

The purpose of the paper is to present a system layout of the new integrated heterogeneous warehousing technology. The present-day enterprises, large and geographically distributed production structures, have accumulated a rapidly increasing, huge data bulk. Currently, the volume of the bulk is measured in petabytes for certain enterprises, and it grows 10-fold every 5 years. Undoubtedly, the warehousing of the bulk is a major issue. The issue is even more challenging in case of a heterogeneous information bulk, which varies from well-structured relational databases to tree and list non-normalized structures, and even to weak-structured multimedia data (scanned papers, audio and video data etc.). The technology outlined below is aimed at more efficient heterogeneous enterprise-level warehousing due to a generalized, unified approach. The technology incorporates a set of models, methods and supporting software tools for object-based data representation and manipulation of heterogeneous enterprise-level warehouses. The technology architecture is based on internet and intranet portals.

2. Heterogeneous Enterprise Warehousing: Problems and Features

Unfortunately, the heterogeneous enterprise warehousing by means of the so-called "industry-level" methodologies (e.g. IBM RUP, Microsoft MSF, Oracle CDM et al.) are not supported by object-based theoretical generalizations, which results in either too narrow spectrum of "mono-vendor" solutions, or to inadequate time-and-cost expenditures. Alternatively, the theoretically promising approaches to data modeling and integration (either category [3], combinatory [4], or ontology-based [7], "SYNTHESIS" project [6] et al.) are too remote from state-of-the-art industry technologies (in CASE and RAD) and thus do not result in software systems with enterprise levels of scalability, expandability, ergonomics etc. A number of federal (Russia, USA, EC) and international (UN, UNESCO), programs recognizing the challenging matter of the issues concerned is another proof of the demand for the new approach to heterogeneous enterprise warehousing.

Thus, the new approach incorporates mathematical models and supporting software toolkit, which provide integration with standard CASE-tools for "industry-level" software development methodologies [9-11]. The approach suggested eliminates data warehouses duplications and contradictions, which essentially increases the enterprise software system (ESS) robustness. The technology addresses a number of interrelated factors of software development, such as programming systems, data models, methodologies and tools, information systems and DBMS architectures [12,13].

The innovative solutions of the technology suggested are:

- General technological outline of ESS development [11];
- Object models for ESS data representation and manipulation [14];
- Problem-oriented visual tools for semantics-based ESS development and content management (ConceptModeller [9], CMIS [12]);
- Portal architecture [13], ESS prototypes and full-scale implementations in a number of enterprises [14].

3. Heterogeneous Enterprise Warehousing: the Integrated Model Set

To adequately model heterogeneous ESS, a systematic approach has been developed, encompassing object models for both data representation and data manipulation.

The general technological outline of ESS development provides a closed-loop, two-way process with reengineering [10,11]. The latter is essential for ESS verification, which significantly increases their robustness.

The new ESS development technology contains stages, corresponding to heterogeneous data representation forms for the globally distributed software components (i.e. natural language, mathematical models, CASE-tools integration, content management etc.) and to the levels, which instantiate these stages (i.e. objects, relationships, events, examples of software toolkits and systems).

The content-oriented approach allows data and metadata generalization on the basis of object models, it also provides for unified manipulation of heterogeneous objects and for adequate internet environment modeling, which is critical for ESS reliability and robustness.

The object nature of the new models is based on "class-object-value" approach, which succeeds both the traditional OOAD and such theoretically promising approaches as (V.E.Wolfengagen's conceptual method [8] and D.Scott's variable domains [2]), and which develops them in the direction of the internet environment [13].

The data model features the following technological transformation sequence:

1. finite sequence term (e.g., λ-calculus term) [1];
2. logical predicate (higher order logic is used) [2];
3. frame (graphical representation) [5];
4. XML object (ConceptModeller class declaration) [9];
5. UML diagram (CASE-integrated data scheme) in ESS repository [10].

Thus, the content representation model is based on semantic networks situation interpretation, which is intuitively transparent for analysts when they chart the problem domain; frame-based visualization provides high ergonomics level of the model. The content management model is based on an abstract machine with states and role assignments, which naturally generalizes the common processes for the similar tools (e.g., building web page templates, publishing web pages, role/access restrictions, etc.). Thus, the major content management operations (such as declaration, evaluation, personalized manipulation etc.) are modeled by the abstract machine language. Syntax and denotational semantics have been produced for the language in terms of variable domains (including content objects construction order, semantics functions and statements for the above operations) [10,13,14].

The data model features the following technological transformation sequence:

1. variable domain term;
2. function over domains (in higher-order logics);
3. frame (graphical representation);
4. XML object (CMS web page template);
5. HTML code on the ESS portal (CMS web page code).

The integrated heterogeneous enterprise warehouse architecture provides high unification degree due to generalized object associations at data and metadata levels. Also, heterogeneous ESS content manipulation is based on a unified internet portal meta-superstructure over the enterprise warehouse. Thus, dynamical, scenario-oriented portal-based content management is provided by assignments, which are implemented as programming language scenarios, switching the abstract machine states.

Scenarios of another type implement personalized content manipulation, supported by a multiparameter functional model and the original Intelligent CMS tool (ICMS) [10-12].

4. The Warehousing Tools Implementation

The ConceptModeller tool provides semantically-oriented visual data scheme development for heterogeneous ESS warehousing. ConceptModeller is based on a semantic network model, which provides intuitive and transparent nearly natural-language support for problem analysts. ConceptModeller warehouse data visualization is based on frame representation of the data scheme [9].

Thus, ConceptModeller provides a self-sufficient, continuous ESS warehouse development cycle from mathematical model to the CASE-level data scheme with reverse engineering. This is due its to integration with the models developed and state-of-the-art CASE-tools (frames are represented by ordered lists).

The ICMS tool is based on an abstract machine model and provides problem-oriented visual manipulation of ESS heterogeneous content and the content publication on the enterprise portal. The ICMS features flexible editing cycle and role mechanisms, which provide content access on the basis of dynamically adaptive profiles and web page templates. Due to scenario oriented content management, the ICMS provides обеспечивает a unified

portal representation of heterogeneous data and metadata objects, flexible content interaction for various user levels (ordinary and privileged, intranet and external), high data security (which is based on access granting scenarios and profiles), a higher ergonomics level (based on flexible personalization) and transparent manipulation of complex data objects (incl. multimedia). Thus, classes are represented by ordered lists of <*attribute, type*>, and templates – by ordered lists of < *attribute, type, value*> [10-12].

5. Conclusion

The approach resulted in a successful implementation of a unified ESS, which integrates such heterogeneous components as state-of-the-art Oracle ERP modules for financial planning and control, a legacy HR system and a weak-structured multimedia warehouse. Internet and intranet portals, manipulating the heterogeneous ESS components, provided a number of implementation in the enterprises of diversified international ITERA Group, incorporating around 10,000 employees in 150 companies of more than 20 countries (http://www.iteragroup.com).

The integral technological approach (mathematical models, SDK tools, portal architecture) provides smooth interaction with a wide range of the state-of-the-art CASE products (IBM Rational, Microsoft Visual Studio .NET, Oracle Developer) and ESS development standards (UML, XML).

Such functional benefits of the approach as compared to the above mentioned competitors as handling complex heterogeneous and variably structured data objects, and integrating architecturally diverse components have become possible – due to model and tool orientation to heterogeneous portal-based ESS. The qualitative assessments of the approach features have been approved by comparison of the major macro indexes (such as TCO, ROI and implementation terms). The ITERA implementation results outperform the leading industry-level methodologies by 30-40% on the average.

Research results-based ESS implementations and curricula have been applied to a number of commercial and governmental enterprise-level organizations (such as ITERA International Group of Companies, Institute for Control Problems of Russian Academy of Sciences, Russian Ministry of Industry and Energy etc.) [13,14].

6. References

[1] Barendregt H.P., 1984. “The lambda calculus” (revised edition), *Studies in Logic*, 103, North Holland, Amsterdam.

[2] Scott D.S., 1982. “Domains for denotational semantics”. In: *ICALP‘82*, pp.577-613.

[3] Cousineau G., Curien P.-L., Mauny M., 1987. “The categorical abstract machine”. *In Science of Computer Programming* 8(2): 173-202.

[4] Curry H.B., Feys R., 1958.*Combinatory logic, Vol.I,* North Holland, Amsterdam.

[5] Roussopulos N.D. “A semantic network model of data bases”, Toronto Univ., 1976.

[6] Kalinichenko L.A., Stupnikov S.A., Martynov D.O. “SYNTHESIS: a Language for Canonical Information Modelling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments”. Moscow: IPI RAN, 2007. - 171 p.

[7] Kleschev A.S., Knyazeva M.A. A concept of the knowledge bank on computer program transformations. *International Journal Information Theories & Applications. , vol. 13, №4,* Bulgaria, Sofia: FOI-COMMERCE. 2006, pp.331-336.

[8] Wolfengagen V.E., 1999. Event-driven objects. In *Proceedings of the 1st International Workshop on Computer Science and information Technologies CSIT’1999.* MEPhI Publishers, Moscow, Russia, Vol.1. p.p. 88-96.

[9] Zykov S.V. “ConceptModeller: Implementing a Semantically-Based Toolkit for Enterprise Applications”. In: *Proc. of the 1st International Conference of Young Scientists on Computer Science and Engineering (CSE-2006)* – Lviv, Ukraine, 2006, pp.23-26.

[10] Zykov S.V. “Enterprise Content Management: Theory and Engineering for Entire Lifecycle Support”. In: *Proc. of the 8th International Workshop on Computer Science and Information Technologies (CSIT'2006)*, Vol. 1, Karlsruhe, Germany, 2006, pp. 86-92.

[11] Zykov S.V. Enterprise Content Management: the Integrated Methodology. In: *B.Granville, N.S.Kutti, M.Missikoff, N.T.Nguyen (Eds.), Enterprise Information Systems and Web Technologies (EISWT’07),* Orlando, FL, U.S.A., July 9-12, 2007, p.p. 226-233.

[12] Zykov S.V. An Integral Approach to Enterprise Content Management. In: *N.Callaos, W.Lesso, C.D.Zinn, B.Zmazek (Eds.), Proc. of International World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2007),* Orlando, FL, U.S.A., July 8-11, 2007, Vol. I, p.p. 212-216.

[13] Zykov S.V. Internet Portal Technologies: Evolution and Examples. In: *Proc. of the 9th International Workshop on Computer Science and Information Technologies (CSIT'2007),* Krasnousolsk, Ufa, Russia, September 13-16, 2007, Vol.1, USATU Editorial-Publishing Office, Ufa, 2007, pp. 99-104.

[14] Zykov S.V. Enterprise Content Management: Bridging the Academia and Industry Gap In: C.Shoniregun, A.Logvinovsky (Eds.), *Proc. of International Conference on Information Society (i-Society 2007),* Merrillville, Indiana, U.S.A., October 7-11, 2007, Vol. I, p.p.145-152.