Yannis Ioannidis
Boris Novikov
Boris Rachev (Eds.)

LNCS 4690

Advances in Databases and Information Systems

11th East European Conference, ADBIS 2007
Varna, Bulgaria, September/October 2007
Proceedings

Springer

Yannis Ioannidis Boris Novikov
Boris Rachev (Eds.)

Advances in Databases and Information Systems

11th East European Conference, ADBIS 2007
Varna, Bulgaria, September 29-October 3, 2007
Proceedings

Springer

Volume Editors

Yannis Ioannidis
University of Athens
Department of Informatics and Telecommunications
Informatics Buildings, Panepistimioupolis, 15784 Ilissia, Athens, Greece
E-mail: yannis@di.uoa.gr

Boris Novikov
University of St.Petersburg
Department of Computer Science
28, Universitetsky Prospekt, Staryj Peterhof, 198504 St.Petersburg, Russia
E-mail: borisnov@acm.org

Boris Rachev
Technical University of Varna
Department of Computer Science and Technologies
1, Studentska Str., 9010 Varna, Bulgaria
E-mail: brachev@gmail.com

Library of Congress Control Number: 2007935199

CR Subject Classification (1998): H.1, H.2, H.3, H.4, H.5

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-75184-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-75184-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12163120 06/3180 5 4 3 2 1 0

Lecture Notes in Computer Science 4690

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

The series of East European Conferences on Advances in Databases and Information Systems (ADBIS) is an established and prestigious forum for the exchange of the latest research results in data management. It provides unique opportunities for database researchers, practitioners, developers, and users from East European countries to explore new ideas, techniques, and tools, and to exchange experiences with colleagues from the rest of the world. This volume contains the proceedings of the 11th ADBIS Conference, held in Varna, Bulgaria, September 29 – October 3, 2007. The conference included 3 keynote talks, 36 research papers in 13 sessions, and 2 tutorials. Twenty-tree of the research papers as well as papers or extended abstracts for the keynote talks are included here; the remaining papers appear in local proceedings.

Distinguished members of the database and information-retrieval communities delivered the three keynotes. Timos Sellis, an expert in the area of multidimensional indexing and data warehousing, analyzed the entire lifecycle of ETL workflows, from specification to optimized execution, offering solutions as well as future challenges. Gerhard Weikum, a leader of several efforts falling at the intersection of databases and information retrieval, discussed the emergence of several "Webs" and how these may be harvested and searched for knowledge. Finally, Paolo Atzeni, well-known for several contributions to database theory, addressed the perennial problem of schema and data translation in the context of emerging model management systems and outlined several research challenges that emerge.

The Research Program Committee consisted of 55 members and was chaired by Yannis Ioannidis (University of Athens, Hellas) and Boris Novikov (University of St. Petersburg, Russia). It accepted 36 papers (23 for the Springer proceedings and 13 for the local proceedings) out of 77 submissions coming from 29 countries. The reviewing process was administrated by the Conference Management System developed and supported by Yordan Kalmukov (University of Rousse). Boris Rachev (Technical University of Varna), Irena Valova (University of Rousse) and Yordan Kalmukov (University of Rousse) edited the proceedings.

The program and social activities of ADBIS 2007 were the result of a huge effort by many hundreds of authors, reviewers, presenters, and organizers. We thank them all for helping to make the conference a success. In particular, we want to thank Peter Antonov (Technical University of Varna) and Angel Smrikarov (University of Rousse) for the smooth local organization.

July 2007

Yannis Ioannidis
Boris Novikov
Boris Rachev

Organization

The 11th East-European Conference on Advances in Databases and Information Systems (ADBIS) was organized by members of the Department of Computer Sciences and Technologies at the Technical University of Varna, and the Department of Computing at the University of Rousse, Bulgaria, in cooperation with the Moscow ACM SIGMOD Chapter.

General Chair

Boris Rachev, Technical University of Varna, Bulgaria

Program Committee Co-chairs

Yannis Ioannidis, National and Kapodistrian University of Athens, Greece
Boris Novikov, University of St. Petersburg, Russia

Program Committee

Antonio Albano (Universita' di Pisa, Italy)
Periklis Andritsos (University of Trento, Italy)
Dmitry Barashev (University of Saint Petersburg, Russia)
Michael Böhlen (Free University of Bozen-Bolzano, Italy)
Stefan Brass (University of Halle, Germany)
Albertas Caplinskas (Institute of Mathematics and Informatics, Lithuania)
Barbara Catania (University of Genoa, Italy)
Chee-Yong Chan (National University of Singapore, Singapore)
Damianos Chatziantoniou (Athens University of Economics and Business, Greece)
Carlo Combi (University of Verona, Italy)
Jens-Peter Dittrich (ETH Zurich, Switzerland)
Pedro Furtado (University of Coimbra, Portugal)
Shahram Ghandeharizadeh (USC, USA)
Maxim Grinev (Institute for System Programming, RAS, Russia)
Oliver Guenther (Humboldt-Universitaet, Germany)
Hele-Mai Haav (Institute of Cybernetics, Estonia)
Piotr Habela (Polish-Japanese Institute of IT, Poland)
Mohand-Said Hacid (University Lyon 1, France)
Theo Härder (University of Kaiserslautern, Germany)
Leonid Kalinichenko (Institute of Informatics Problems, RAS, Russia)
Mikhail Kogalovsky (Market Economy Institute of RUS, Russia)
Georgia Koutrika (Stanford University, USA)

Sergei Kuznetsov (Institute for System Programming, Russia)
Nikos Mamoulis (University of Hong Kong, Hong Kong)
Yannis Manolopoulos (Aristotle University, Greece)
Rainer Manthey (University of Bonn, Germany)
Tadeusz Morzy (Institute of Computing Science, Poland)
Thomas Neumann (Max-Planck-Institut für Informatik, Germany)
Kjetil Norvag (Norwegian University of Science and Technologies, Norway)
Fatma Ozcan (IBM Almaden Research Center, USA)
Jaroslav Pokorny (Charles University, Czech Republic)
Alkis Polyzotis (University of California, Santa Cruz, USA)
Alexandra Poulovassilis (Birkbeck, University of London, UK)
Manfred Reichert (University of Twente, Netherlands)
Karel Richta (Czech Technical University, Czech Republic)
George Samaras (University of Cyprus, Cyprus)
Peter Scheuermann (Northwestern University, USA)
Heiko Schuldt (Database and Information Systems Group, Switzerland)
Dmitry Shaporenkov (University of Saint Petersburg, Russia)
Serge Shumilov (University of Bonn, Germany)
Alkis Simitsis (IBM Almaden Research Center, USA)
Leonid Sokolinsky (Southern Ural State University, Russia)
Bernhard Thalheim (CAU Kiel, Germany)
Martin Theobald (Stanford University, USA)
Yannis Theodoridis (University of Piraeus, Greece)
Ricardo Torlone (Università Roma Tre, Italy)
Christos Tryfonopoulos (Max-Planck-Institut für Informatik, Germany)
Toni Urpí (UPC, Spain)
Panos Vassiliadis (University of Ioannina, Greece)
Jari Veijalainen (University of Jyvaskyla, Finland)
Yannis Velegrakis (University of Trento, Italy)
Stratis Viglas (University of Edinburgh, UK)
Marek Wojciechowski (Poznan University of Technology, Poland)
Robert Wrembel (Poznan University of Technology, Poland)
Vladimir Zadorozhny (University of Pittsburgh, USA)

Additional Reviewers

Mehmet Altinel
Ralph Bobrik
Antonio Corral
Sergei Evdokimov
Paul El Khoury
Maria Kontaki
Lyes Liman
Amine Mokhtari
Alexandros Nanopoulos
René Noack
George Pallis
Paraskevi Raftopoulou
Ramzi Rizk
Salvatore Ruggieri
Carlo Sartiani
Angela Siu
Goce Trajcevski
Man Lung Yiu
Thomas Zahn

Organizing Committee Chairs

Peter Antonov, Technical University of Varna
Angel Smrikarov, University of Rousse

Organizing Committee Coordinators

Irena Valova, University of Rousse
Yordan Kalmukov, University of Rousse

Organizing Committee Members

Milko Marinov, University of Rousse
Yulka Petkova, Technical University of Varna
Silyan Arsov, University of Rousse
Antoaneta Ivanova, Technical University of Varna
Veneta Aleksieva, Technical University of Varna
Miroslava Strateva, Technical University of Varna
Polina Kalmukova, Graphic Designer

ADBIS Steering Committee Chair

Leonid Kalinichenko, Russian Academy of Science, Russia

ADBIS Steering Committee

Andras Benczur (Hungary)
Albertas Caplinskas (Lithuania)
Johann Eder (Austria)
Janis Eiduks (Latvia)
Hele-Mai Haav (Estonia)
Mirjana Ivanovic (Serbia)
Mikhail Kogalovsky (Russia)
Yannis Manolopoulos (Greece)
Rainer Manthey (Germany)
Tadeusz Morzy (Poland)
Pavol Navrat (Slovakia)
Boris Novikov (Russia)
Jaroslav Pokorny (Czech Republic)
Boris Rachev (Bulgaria)
Anatoly Stogny (Ukraine)
Bernhard Thalheim (Germany)
Tatjana Welzer (Slovenia)
Viacheslav Wolfengagen (Russia)

Lecture Notes in Computer Science

Sublibrary 3: Information Systems and Application, incl. Internet/Web and HCI

For information about Vols. 1– 4277
please contact your bookseller or Springer

Vol. 4740: L. Ma, M. Rauterberg, R. Nakatsu (Eds.), Entertainment Computing – ICEC 2007. XXX, 480 pages. 2007.

Vol. 4730: C. Peters, P. Clough, F.C. Gey, J. Karlgren, B. Magnini, D.W. Oard, M. de Rijke, M. Stempfhuber (Eds.), Evaluation of Multilingual and Multi-modal Information Retrieval. XXIV, 998 pages. 2007.

Vol. 4723: M. R. Berthold, J. Shawe-Taylor, N. Lavrač (Eds.), Advances in Intelligent Data Analysis VII. XIV, 380 pages. 2007.

Vol. 4718: J. Hightower, B. Schiele, T. Strang (Eds.), Location- and Context-Awareness. X, 297 pages. 2007.

Vol. 4717: J. Krumm, G.D. Abowd, A. Seneviratne, T. Strang (Eds.), UbiComp 2007: Ubiquitous Computing. XIX, 520 pages. 2007.

Vol. 4715: J.M. Haake, S.F. Ochoa, A. Cechich (Eds.), Groupware: Design, Implementation, and Use. XIII, 355 pages. 2007.

Vol. 4690: Y. Ioannidis, B. Novikov, B. Rachev (Eds.), Advances in Databases and Information Systems. XIII, 377 pages. 2007.

Vol. 4675: L. Kovács, N. Fuhr, C. Meghini (Eds.), Research and Advanced Technology for Digital Libraries. XVII, 585 pages. 2007.

Vol. 4674: Y. Luo (Ed.), Cooperative Design, Visualization, and Engineering. XIII, 431 pages. 2007.

Vol. 4663: C. Baranauskas, P. Palanque, J. Abascal, S.D.J. Barbosa (Eds.), Human-Computer Interaction – INTERACT 2007, Part II. XXXIII, 735 pages. 2007.

Vol. 4662: C. Baranauskas, P. Palanque, J. Abascal, S.D.J. Barbosa (Eds.), Human-Computer Interaction – INTERACT 2007, Part I. XXXIII, 637 pages. 2007.

Vol. 4658: T. Enokido, L. Barolli, M. Takizawa (Eds.), Network-Based Information Systems. XIII, 544 pages. 2007.

Vol. 4656: M.A. Wimmer, J. Scholl, Å. Grönlund (Eds.), Electronic Government. XIV, 450 pages. 2007.

Vol. 4655: G. Psaila, R. Wagner (Eds.), E-Commerce and Web Technologies. VII, 229 pages. 2007.

Vol. 4654: I.Y. Song, J. Eder, T.M. Nguyen (Eds.), Data Warehousing and Knowledge Discovery. XVI, 482 pages. 2007.

Vol. 4653: R. Wagner, N. Revell, G. Pernul (Eds.), Database and Expert Systems Applications. XXII, 907 pages. 2007.

Vol. 4636: G. Antoniou, U. Aßmann, C. Baroglio, S. Decker, N. Henze, P.-L. Patranjan, R. Tolksdorf (Eds.), Reasoning Web. IX, 345 pages. 2007.

Vol. 4611: J. Indulska, J. Ma, L.T. Yang, T. Ungerer, J. Cao (Eds.), Ubiquitous Intelligence and Computing. XXIII, 1257 pages. 2007.

Vol. 4607: L. Baresi, P. Fraternali, G.-J. Houben (Eds.), Web Engineering. XVI, 576 pages. 2007.

Vol. 4606: A. Pras, M. van Sinderen (Eds.), Dependable and Adaptable Networks and Services. XIV, 149 pages. 2007.

Vol. 4605: D. Papadias, D. Zhang, G. Kollios (Eds.), Advances in Spatial and Temporal Databases. X, 479 pages. 2007.

Vol. 4602: S. Barker, G.-J. Ahn (Eds.), Data and Applications Security XXI. X, 291 pages. 2007.

Vol. 4592: Z. Kedad, N. Lammari, E. Métais, F. Meziane, Y. Rezgui (Eds.), Natural Language Processing and Information Systems. XIV, 442 pages. 2007.

Vol. 4587: R. Cooper, J. Kennedy (Eds.), Data Management. XIII, 259 pages. 2007.

Vol. 4577: N. Sebe, Y. Liu, Y.-t. Zhuang, T.S. Huang (Eds.), Multimedia Content Analysis and Mining. XIII, 513 pages. 2007.

Vol. 4568: T. Ishida, S. R. Fussell, P. T. J. M. Vossen (Eds.), Intercultural Collaboration. XIII, 395 pages. 2007.

Vol. 4566: M.J. Dainoff (Ed.), Ergonomics and Health Aspects of Work with Computers. XVIII, 390 pages. 2007.

Vol. 4564: D. Schuler (Ed.), Online Communities and Social Computing. XVII, 520 pages. 2007.

Vol. 4563: R. Shumaker (Ed.), Virtual Reality. XXII, 762 pages. 2007.

Vol. 4561: V.G. Duffy (Ed.), Digital Human Modeling. XXIII, 1068 pages. 2007.

Vol. 4560: N. Aykin (Ed.), Usability and Internationalization, Part II. XVIII, 576 pages. 2007.

Vol. 4559: N. Aykin (Ed.), Usability and Internationalization, Part I. XVIII, 661 pages. 2007.

Vol. 4558: M.J. Smith, G. Salvendy (Eds.), Human Interface and the Management of Information, Part II. XXIII, 1162 pages. 2007.

Vol. 4557: M.J. Smith, G. Salvendy (Eds.), Human Interface and the Management of Information, Part I. XXII, 1030 pages. 2007.

Vol. 4541: T. Okadome, T. Yamazaki, M. Makhtari (Eds.), Pervasive Computing for Quality of Life Enhancement. IX, 248 pages. 2007.

Vol. 4537: K.C.-C. Chang, W. Wang, L. Chen, C.A. Ellis, C.-H. Hsu, A.C. Tsoi, H. Wang (Eds.), Advances in Web and Network Technologies, and Information Management. XXIII, 707 pages. 2007.

Vol. 4531: J. Indulska, K. Raymond (Eds.), Distributed Applications and Interoperable Systems. XI, 337 pages. 2007.

Vol. 4526: M. Malek, M. Reitenspieß, A. van Moorsel (Eds.), Service Availability. X, 155 pages. 2007.

Vol. 4524: M. Marchiori, J.Z. Pan, C.d.S. Marie (Eds.), Web Reasoning and Rule Systems. XI, 382 pages. 2007.

Vol. 4519: E. Franconi, M. Kifer, W. May (Eds.), The Semantic Web: Research and Applications. XVIII, 830 pages. 2007.

Vol. 4518: N. Fuhr, M. Lalmas, A. Trotman (Eds.), Comparative Evaluation of XML Information Retrieval Systems. XII, 554 pages. 2007.

Vol. 4508: M.-Y. Kao, X.-Y. Li (Eds.), Algorithmic Aspects in Information and Management. VIII, 428 pages. 2007.

Vol. 4506: D. Zeng, I. Gotham, K. Komatsu, C. Lynch, M. Thurmond, D. Madigan, B. Lober, J. Kvach, H. Chen (Eds.), Intelligence and Security Informatics: Biosurveillance. XI, 234 pages. 2007.

Vol. 4505: G. Dong, X. Lin, W. Wang, Y. Yang, J.X. Yu (Eds.), Advances in Data and Web Management. XXII, 896 pages. 2007.

Vol. 4504: J. Huang, R. Kowalczyk, Z. Maamar, D. Martin, I. Müller, S. Stoutenburg, K.P. Sycara (Eds.), Service-Oriented Computing: Agents, Semantics, and Engineering. X, 175 pages. 2007.

Vol. 4500: N.A. Streitz, A. Kameas, I. Mavrommati (Eds.), The Disappearing Computer. XVIII, 304 pages. 2007.

Vol. 4495: J. Krogstie, A. Opdahl, G. Sindre (Eds.), Advanced Information Systems Engineering. XVI, 606 pages. 2007.

Vol. 4480: A. LaMarca, M. Langheinrich, K.N. Truong (Eds.), Pervasive Computing. XIII, 369 pages. 2007.

Vol. 4471: P. Cesar, K. Chorianopoulos, J.F. Jensen (Eds.), Interactive TV: A Shared Experience. XIII, 236 pages. 2007.

Vol. 4469: K.-c. Hui, Z. Pan, R.C.-k. Chung, C.C.L. Wang, X. Jin, S. Göbel, E.C.-L. Li (Eds.), Technologies for E-Learning and Digital Entertainment. XVIII, 974 pages. 2007.

Vol. 4443: R. Kotagiri, P. Radha Krishna, M. Mohania, E. Nantajeewarawat (Eds.), Advances in Databases: Concepts, Systems and Applications. XXI, 1126 pages. 2007.

Vol. 4439: W. Abramowicz (Ed.), Business Information Systems. XV, 654 pages. 2007.

Vol. 4430: C.C. Yang, D. Zeng, M. Chau, K. Chang, Q. Yang, X. Cheng, J. Wang, F.-Y. Wang, H. Chen (Eds.), Intelligence and Security Informatics. XII, 330 pages. 2007.

Vol. 4425: G. Amati, C. Carpineto, G. Romano (Eds.), Advances in Information Retrieval. XIX, 759 pages. 2007.

Vol. 4412: F. Stajano, H.J. Kim, J.-S. Chae, S.-D. Kim (Eds.), Ubiquitous Convergence Technology. XI, 302 pages. 2007.

Vol. 4402: W. Shen, J.-Z. Luo, Z. Lin, J.-P.A. Barthès, Q. Hao (Eds.), Computer Supported Cooperative Work in Design III. XV, 763 pages. 2007.

Vol. 4398: S. Marchand-Maillet, E. Bruno, A. Nürnberger, M. Detyniecki (Eds.), Adaptive Multimedia Retrieval: User, Context, and Feedback. XI, 269 pages. 2007.

Vol. 4397: C. Stephanidis, M. Pieper (Eds.), Universal Access in Ambient Intelligence Environments. XV, 467 pages. 2007.

Vol. 4380: S. Spaccapietra, P. Atzeni, F. Fages, M.-S. Hacid, M. Kifer, J. Mylopoulos, B. Pernici, P. Shvaiko, J. Trujillo, I. Zaihrayeu (Eds.), Journal on Data Semantics VIII. XV, 219 pages. 2007.

Vol. 4365: C.J. Bussler, M. Castellanos, U. Dayal, S. Navathe (Eds.), Business Intelligence for the Real-Time Enterprises. IX, 157 pages. 2007.

Vol. 4353: T. Schwentick, D. Suciu (Eds.), Database Theory – ICDT 2007. XI, 419 pages. 2006.

Vol. 4352: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part II. XVIII, 743 pages. 2006.

Vol. 4351: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part I. XIX, 797 pages. 2006.

Vol. 4328: D. Penkler, M. Reitenspiess, F. Tam (Eds.), Service Availability. X, 289 pages. 2006.

Vol. 4321: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), The Adaptive Web. XII, 763 pages. 2007.

Vol. 4317: S.K. Madria, K.T. Claypool, R. Kannan, P. Uppuluri, M.M. Gore (Eds.), Distributed Computing and Internet Technology. XIX, 466 pages. 2006.

Vol. 4312: S. Sugimoto, J. Hunter, A. Rauber, A. Morishima (Eds.), Digital Libraries: Achievements, Challenges and Opportunities. XVIII, 571 pages. 2006.

Vol. 4306: Y. Avrithis, Y. Kompatsiaris, S. Staab, N.E. O'Connor (Eds.), Semantic Multimedia. XII, 241 pages. 2006.

Vol. 4302: J. Domingo-Ferrer, L. Franconi (Eds.), Privacy in Statistical Databases. XI, 383 pages. 2006.

Vol. 4299: S. Renals, S. Bengio, J.G. Fiscus (Eds.), Machine Learning for Multimodal Interaction. XII, 470 pages. 2006.

Vol. 4295: J.D. Carswell, T. Tezuka (Eds.), Web and Wireless Geographical Information Systems. XI, 269 pages. 2006.

Vol. 4286: P.G. Spirakis, M. Mavronicolas, S.C. Kontogiannis (Eds.), Internet and Network Economics. XI, 401 pages. 2006.

Vol. 4282: Z. Pan, A. Cheok, M. Haller, R.W.H. Lau, H. Saito, R. Liang (Eds.), Advances in Artificial Reality and Tele-Existence. XXIII, 1347 pages. 2006.

Vol. 4278: R. Meersman, Z. Tari, P. Herrero (Eds.), On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Part II. XLV, 1004 pages. 2006.

Table of Contents

Object-Oriented Systems

Indexing

Clustering and OLAP

Moving Objects

Query Processing

DB Architectures and Streams

XML and Databases

Distributed Systems

ETL Workflows: From Formal Specification to Optimization

Timos K. Sellis and Alkis Simitsis

[1] School of Electrical and Computer Engineering,
National Technical University of Athens, Athens, Hellas
timos@dblab.ece.ntua.gr
[2] IBM Almaden Research Center, San Jose CA 95120, USA
asimits@us.ibm.com

Abstract. In this paper, we present our work on a framework towards the modeling and optimization of Extraction-Transformation-Loading (ETL) workflows. The goal of this research was to facilitate, manage, and optimize the design and implementation of the ETL workflows both during the initial design and deployment stage, as well as, during the continuous evolution of a data warehouse. In particular, we present our results which include: (a) the provision of a novel conceptual model for the tracing of inter-attribute relationships and the respective ETL transformations in the early stages of a data warehouse project, along with an attempt to use ontology-based mechanisms to semi-automatically capture the semantics and the relationships among the various sources; (b) the provision of a novel logical model for the representation of ETL workflows with two main characteristics: genericity and customization; (c) the semi-automatic transition from the conceptual to the logical model for ETL workflows; and (d) the tuning of an ETL workflow for the optimization of the execution order of its operations. Finally, we discuss some issues on future work in the area that we consider important and a step towards the incorporation of the above research results to other areas as well.

1 Introduction

Successful planning and decision making in large enterprises requires the ability of efficiently processing and analyzing the organization's informational data. Such data is typically distributed in several heterogeneous sources and stored under different structures and formats. To deal with such issues, as well as performance issues, and to support the functionality of On Line Analytical Processing (OLAP) applications and Decision Support Systems (DSS), Data Warehouses (DW) are employed to integrate the data and provide an appropriate infrastructure for querying, reporting, mining, and other advanced analysis techniques. The procedure of designing and populating a DW has been characterized as a very troublesome and time consuming task with a significant cost in human, system, and financial resources [13].

In the past, research has treated DW as collections of materialized views. Although this abstraction may suffice for the purpose of examining alternative strategies for view maintenance, it can not adequately describe the structure and contents of a DW in real-world settings. A more elaborated approach is needed (a) to represent the population of

Y. Ioannidis, B. Novikov, and B. Rachev (Eds.): ADBIS 2007, LNCS 4690, pp. 1–11, 2007.

the DW with data stemming from a set of heterogeneous source datastores, and (b) to take into consideration that during their transportation, data may be transformed to meet the schema and business requirements of the DW. This procedure normally composes a labor intensive workflow and constitutes an integral part of the back-stage of DW architectures.

Hence, to overcome the above problems, specialized workflows are used under the general title *Extraction-Transformation-Loading* (ETL) workflows. ETL workflows represent an important part of data warehousing, as they represent the means in which data actually gets loaded into the warehouse. Their generic functionality and most prominent tasks include:

- the identification of relevant information at the source side,
- the extraction of this information,
- the transportation of this information to the Data Staging Area (DSA), where, usually, all transformations take place,
- the transformation, (i.e., customization and integration) of the information coming from multiple sources into a common format,
- the cleansing of the resulting data set, on the basis of database and business rules, and
- the propagation and loading of the data to the DW and the refreshment of data marts.

Figure 1 depicts a generic architecture of the DW environment.

Several research approaches have studied the modeling part of ETL processes. On the other hand, several commercial tools already exist in the market and all major DBMS vendors provide such functionality. However, each individual effort follows a different approach for the modeling and representation of ETL processes, making essential the adoption of an unified formal description of such processes. For a further discussion on the importance of ETL processes and on the problems existing due to the lack of a uniform modeling technique, along with a review of the state of the art in both research and commercial solutions, we refer the interested reader to [7].

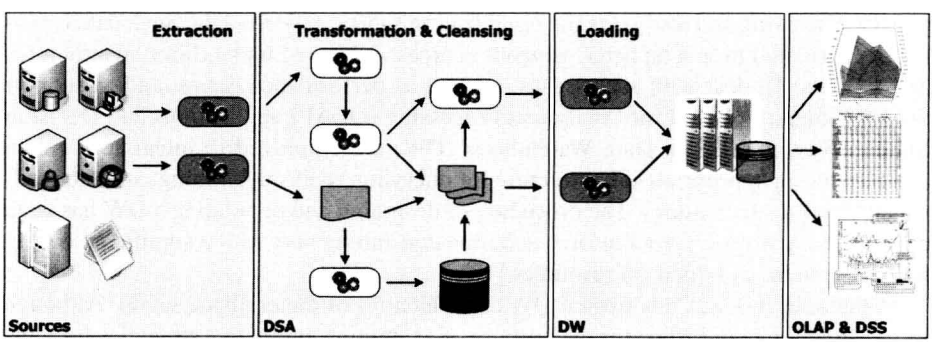

Fig. 1. Generic Architecture of Data Warehouse environment

In this paper, we present our work towards the modeling and optimization of ETL workflows. Section 2 presents our framework for the formal specification of ETL workflows. Section 3 describes our technique for the logical optimization of ETL workflows. Finally, Section 4 concludes our discussion with a prospect of the future.

2 Formal Specification of ETL Workflows

2.1 Identification of ETL Design Requirements Using Semantic Web Technology

During the initial phases of a DW design and deployment, one of the main challenges is the identification of the involved sources and the determination of appropriate inter-schema mappings and transformations from the data sources to the DW. Currently, research and commercial ETL tools mainly focus on the representation and design of ETL scenarios. The identification of the required mappings and transformations is done manually, due to the lack of precise metadata, regarding the semantics of the data sources and the constraints and requirements of the DW. Hence, such information is incomplete or even inconsistent, often being hard-coded within the schemata or metadata of the sources or even provided in natural language format after oral communication with the involved parties; e.g., business managers and administrators/designers of the DW. As a result, designing ETL processes becomes a very tedious and error-prone task. Given the fact that typical ETL processes are quite complex and that significant operational problems can occur with improperly designed ETL systems, developing a formal, metadata-driven approach to allow a high degree of automation of the ETL design, is critical in employing a Data Warehouse solution.

In our research, we have worked on the aforementioned problem. Earlier work argues that in the context of a DW application, ontologies constitute a suitable conceptual model for describing the semantics of the datastores and automatically identifying relationships among them using reasoning techniques [9,11]. The schema of a datastore describes the way that data is structured when stored, but does not provide any information for its intended semantics. Therefore, metadata are required to allow for the understanding, management, and processing of this data. Semantic Web technologies provide a means to formally specify the metadata, so that automated reasoning techniques can be used to facilitate further processing.

A graph-based representation, called *datastore graph*, is employed as a common model for the datastores. Graphs constitute a generic data model allowing the representation of several types of schemas, including relational and XML schemas, thereby allowing for both structured and semi-structured sources to be handled in a unified way. A graph representation, termed ontology graph, is introduced for the application ontology. Providing a visual, graph-based representation, with different symbols for the different types of classes and properties in the ontology, makes it easier for the designer to create, verify and maintain the ontology, as well as use it as a means of communication between different parties involved in the project.

Annotation of a datastore is accomplished by formally defining mappings between the nodes of the datastore graph and the ontology graph. These mappings can be represented as labels assigned to the nodes of the data store graph, i.e., the datastore is

semantically described by the annotated datastore graph. The mappings may be specified either (semi-)automatically using results provided by related research efforts [2] or manually – e.g., by implementing drag-and-drop functionality between the visual representations of the corresponding graphs. In both cases, the time and effort required for establishing and maintaining the mappings significantly decreases with respect to common practice.

Based on the application ontology and the annotated datastore graphs, automated reasoning techniques are used to infer correspondences and conflicts among the datastores, thus, identifying relevant sources and proposing conceptual operations for integrating data into the DW.

Furthermore, the application ontology along with a common application terminology, can be used as a common language, to produce a textual description of the requirements of an ETL process. The verbalization of such requirements further facilitates the communication among the involved parties and the overall process of design, implementation, maintenance, and documentation. Recent results describe how a common application terminology can be established semi-automatically, using linguistic techniques [10]. In that work, a template-based technique is introduced to represent the semantics and the metadata of ETL processes as a narrative, based on information stored in the application ontology, which captures business requirements, documentation, and existing schemata. In addition, the customization and tailoring of reports to meet diverse information needs, as well as the grouping of related information to produce more concise and comprehensive output are considered.

The result of the above work is accompanied by a simple graphical model, which facilitates the smooth redefinition and revision efforts and serves as the means of communication with the rest of the involved parties [13]. A graph-based representation of the involved datastores and transformations is presented in a customizable and extensible manner. The transformations used in this model follow a high level description annotated with sufficient information for their ensuing formal specification in the logical level. (For a further analysis on this issue, we defer to subsection 2.3.)

2.2 Logical Modeling of ETL Workflows

A conceptual model for ETL processes serves as a suitable means for communications and requirements understanding in the early stages of a DW project during which, the time constraints of the project require a quick documentation of the involved data stores and their relationships, rather than an in-depth description of a composite workflow. For the ensuing stages of the project, a formal and more rigorous logical model is necessary.

In our research, we have extensively dealt with this challenge by presenting a formal *logical model* specifically tailored for the ETL environment [8,12,14]. The model concentrates on the flow of data from the sources towards the data warehouse through the composition of activities (transformations) and datastores. The core of the proposed model treats an ETL scenario as a graph of ETL activities having interconnected input and output schemata. This graph, which is referred to as Architecture Graph, can be used as the blueprints for the structure of an appropriate workflow in repository management, visualization, and what-if analysis tools. Activities, datastores, and their respective attributes are modeled as the nodes of the graph. Provider relationships that