File and database integrated databse transformations: An middleware

Sudharsun. P. R, Pradeep. V. Saravana Kumar. R.

Abstract— The database transformations have either been performed using 'Command line interface' or import tools. This traditional process is complex and becomes tedious when same set of transformation has to be performed periodically. Moreover, Data transformations are not supported for varying schemas and outdated database applications. The goal of our project is to develop a 'middleware application' which eases the task of database transformations over a vast range of databases and flat file structures, and varying schemas.

Index Terms— Database Transformations, ETL process, Middleware application, Import/Export database, Database, Database Application.

1 INTRODUCTION

DATABASE transformation is the process of transferring a table or th whole schema of a database, from one database to another. This process may involve changing the structure, type, format, definition and (or) involes changing from one database application to another. Earlier this process is commonly seen in data mining and warehousing systems, but now when data has growm larger and huge it is getting familier in databases too.

Earlier data base transformations were performaed using command line interface or import/ Export tools. At that time data was small and hence it was easier to perform this process via comman line interface and (or) import/ export tools. Now since the data is become huge, it becomes a harazzing job to perform the transformation using traditional methods and require automaion systems.

Database Transformation is a 3 stage process namely, Extract, Transform, Load. (ETL). In the Extract procee, the data from the database source is obtained. In the Transform process, the data undergoes all necessary transformation such as structure, format type, definition, etc. In the Load process, the transformed data is stored into the Target database

- *Mr. P. R. Sudharsun. is currently pursuing bachelor's degree program in computer science engineering in GKM college of Enginerring and Technology, India, PH-+918056027716. E-mail: sudharsunpr@live.in*
- *Mr. V. Pradeep is currently pursuing bachelor's degree program in computer science engineering in GKM college of Enginerring and Technology, India, PH-+919445324488. E-mail: pradeepleoa175@gmail.com*
- *Mr.R. Saravana Kumar. is currently pursuing bachelor's degree program in comuter science engineering in GKM college of Enginerring and Technology, India, PH-+919789098090. E-mail: r.saravanakumar10@yahoo.com*

2 EXSISTING SYSTEM

The existing system allowes database transformations to be performed either using command line interface or import / Export tools. It does not support many to one transformations. The traditional process has a high memory and data overload since the import and export have to be performed separately, when done via command line interface. The traditional process becomes a hectic one, which does not have featires such as configuration profiling, automation, logging, polling concept, bagroung threading, etc.

The existing system does not allow transformation to be performed on many to one architecture at a single instance. And majour disadvantage found during our research wad memory overload while transforming from one database application to another application using command line interface, since both the database instances nedd to be running.

3 LITERATURE SURVEY

3.1 Research on Heterogeneous Database Transformation

This paper appears in:
Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference
Publisher: IEEE Computer Society Washington, DC, USA ©2011
Date of Conference: 28-29 March 2011
Author(s): Shudao, Zhou
Inst. of Meteorol., PLA Univ. of Sci. & Tech., Nanjing, China
Guotao, Zhu; Yanjie, Wang; Ye, Shen; Zanyi, Liu
Abstract: Access, MySQL, SQL Server and Oracle9i databases are studyed in this paper. Aiming at the characteristics of data types in heterogeneous database, simple but flexible data types matching method is proposed. In the environment of C#.Net, using standard SQL language and binary text file as intermediate data file, heterogeneous database import-and-export system is developped, which is relatively independent to individual DBMSs. The practical application results show that the method is rapid and feasible on data exchange for heterogeneous databases.

From this paper we find that using a flat file or a binary file

as intermediate for heterogenous database transformation the perfomaces such as memory, speed, disk usage, etc,. have been drastically imporoved

We aso find that using XML database as a intermediate in Java environment the perfoemance of the transformation middlware has imporver by 2x factor.

3.2 Automated Structure Extraction and XML Conversion of Life Science Database Flat Files

This paper appears in:
Information Technology in Biomedicine, IEEE Transactions
Publisher: IEEE Computer Society Washington, DC, USA
©2011
Date of Conference: October 2006
Author(s): Philippi, Stephan
Univ. of Koblenz
Köhler, Jacob
Abstract: In the light of the increasing number of biological databases, their integration is a fundamental prerequisite for answering complex biological questions. Database integration, therefore, is an important area of research in bioinformatics. Since most of the publicly available life science databases are still exclusively exchanged by means of proprietary flat files, database integration requires parsers for very different flat file formats. Unfortunately, the development and maintenance of database specific flat file parsers is a nontrivial and time-consuming task, which takes considerable effort in large-scale integration scenarios. This paper introduces heuristically based concepts for automatic structure extraction from life science database flat files. On the basis of these concepts the FlatEx prototype is developed for the automatic conversion of flat files into XML representations.

From this paper, since we are using XML as an intermediate for heterogenous database transformation we get the concepts of XML parsing and conversion of heterogenous databases.

3.3 Transformation of Flat File into Data Warehouse

This paper appears in:
Global Journal of Computer Science and Technology
Publisher: Global Journals Inc. (USA)
Date of Conference: Augest 2011.
Author(s): By Muhammad Inayat Ullah, Muhammad Zeeshan, Mahwish Kundi, omal University,Dera Ismail Khan, Pakistan.
Abstract: A Flat file (Semi Structured) Data comes from different sources or operational systems for storage in the data warehouse. Extraction, transformation and loading of the data could be necessary. Moreover, input flat file data must be transformed into a uniform format which could be more suitable for analytical purposes. Aim of this research is to analysis the delimiters of the flat file, to transform flat file into uniform format and suggest a suitable algorithm for implementation such type of algorithm could be solve the problem of transformation of the flat file data and such algorithm could be useful for extraction, transformation and loading of huge amount of flat file data into data warehouse.

From this paper, since we are using XML as a intermediate to transformation in our middlware we gracps the concepts and technology used in this paper to transform a flat file into another heterogenous database

3.4 Signature search time evaluation in Flat file databases

This paper appears in:
Aerospace and Elctronic Systems, IEEE transactions
Publisher: IEEE Computer Society Washington, DC, USA
Date of Conference: April 2008.
Author(s): Ko, Kwangil
Stony Brook Univ., Stony Brook, NY
Robertazzi, Thomas G.
Abstract: For the first time, divisible load scheduling theory is used to solve for the expected time for searching for both single and multiple signatures in certain multiple processor database architectures. The target architectures examined for illustrative purposes are linear daisy chains and single level tree networks with single and multiple installment load distribution. The use of divisible load modeling and analysis yields elegant expressions for expected search time.

From this paper, we grasps the concepts of searching and evaluvating the flat file database since our middleware supports database transformation on Flat File databases. Flat file database can be a plain text file or binary file where single fields can be separated by delimeiters such as comma, special characters.

3.5 Transforming Heterogeneous Data with Database Middleware: Beyond Integration

This paper appears in:
Technical Committee on Data Engineering, IEEE transactions
Publisher: IEEE Computer Society Washington, DC, USA
Date of Conference: 1997.
Author(s): L. M. Haas R. J. Miller B. Niswonger M. Tork Roth P. M. Schwarz E. L. Wimmers
Abstract: Many applications today need information from diverse data sources, in which related data may be represented quite differently. In one common scenario, a DBA wants to add data from a new source to an existing warehouse. The data in the new source may not match the existing warehouse schema. The new data may also be partially redundant with that in the existing warehouse, or formatted differently. Other applications may need to integrate data more dynamically, in response to user queries. Even applications using data from a single source often want to present it in a form other than that it is stored in. This progect proposes a Database Midlleware.

This paper serves as the base paper of this project with an imporved architectureal design to improve the performance of both the client system and the middleware, with XML as an intermediate. From this paper, we study the concept of integrating the heterogenous database to perform ETL process. Our research suggests that the results are drastically improved in many aspects.

3 PROPOSED SYSTEM

The ultimate aim of this project is to develop a middlware system that suppost heterogenous database transformation over a vast range of Enterprise database applications such as Access, Oracel 9i, 10g 11i, Oracle MySQL, MS Access, MS Excel, IBM DB2, etc,. And flat file database in a Java Environment and using XMl as an improvement. Additional Features include Confguation profiling, Polling interval, eliminating delimiter and column padding along with logging.

4 ARCHITECTURE OF THE MIDDLEWARE SYSTEM

This project poposes a fused schema which intergrates vast range of heterogenous databases that supports database transformations over database and flat files using XML as an intermediate.

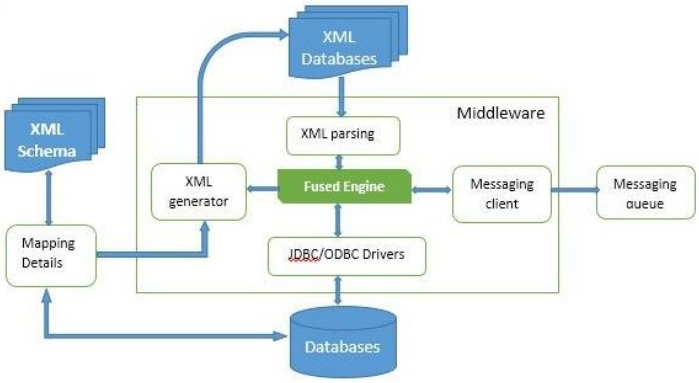

Figure 1. Architecture of the Middleware system.

The architecture of the Middleware can be developed in two versions namely: Standalone application with java database and application with Enterprise database application for configuration profile storage. Each of which has its own features which is compared in the table below.

TABLE 1
COMPARITION BETWEEN MIDDLEWARE VERSIONS

Feature	Standalone application	Enterprise database application
Support for:		
Flat files	Yes	Yes
Oralcle	Yes	Yes
My SQL	Yes	Yes
Database details visibleity	No	Yes
Password encryption	No Feature	Yes
Tranceparency	Yes	Yes
Time*	1.775ms	2ms
Ram Usage*	2%	10%
Accuracy	Yes	Yes
XML Input	Yes	Yes
Background running	Yes	No

*Time is for transformation between hetrogenous databases with 5 fields with 1000 rows and Ram Usage is at the time of transformation.

4.1 Tool algorithm

The Algorithm of the tool working is as follows

1 Start
2 Get and save the source and target database field details.
3 Get and save mapping detatils.
4 Get and save the type of source and target database
5 Get and save the configuration of the source ans target database.
6 Get the polling interval
7 Save the configuration
8 Run the ETL process
   a. Loop for poling interval upto time
      i. If source is a flat file check if file details and format is correct.
      ii. If source is a database check database details and configuration.
      iii. If errors were found inform the client via message queue.
      iv. Generate source database XML with source design and configuration details.
      v. Generate database XMl with the help of Source XML and Mapping Data.
      vi. If Target is a flat file, check if file is present. If no create a file.
      vii. If target is a database check if table/ schema is present.
      viii. If errors were found inform the client via message queue.
      ix. Insert target data with the help of target XML via XML parsing in Java environment.
      x. If Target is a flat file Remove rows based on primary key
      xi. Similar cases apply id target is a database.
   b. End loop
   c. Log the details of ETL.
9 End background threads.
10 Run Garbage Collector.
11 Exit upon Client reques.

5 FUTURE ENHANCEMET

Since Data source are not just from Standalone database applications and flat file, this middleware is not sufficient to integrae and transform data from web services and service oriented architecture. This paper propose a middlware application in th future to perform database transformation over service oriented architecture and web services along with the exsistent

support on file and standalone enterprise databases.

6 FUTURE ENHANCEMET

Thus the final database can integrate and perform database transformation over various sources of enterprise database application and flat files with higher security and accuracy. The database transformation performed is much more transparent, scalable and high performed in various aspects such as time, RAM & disk usage, etc, making the middlware a successful one.

ACKNOWLEDGMENT

Firstly we would like to thank our institution, GKM college of Engineering and Technology, Chennai, India, for providing us with various services such as internet, R&D department, IEEE and other journal access, etc. which helped us developing the project. We also wish to thank our head of the Department, Ms. Neelavani and Ms. Sudha Assistant professer, CSE for extended support for making this project a successful one. We also extend our thank to all the staff me,ber of our department for raising queries at time to time and supporting us to make the project a successful one.

REFERENCES

[1] Muhammad Inayat Ullah, Muhammad Zeeshan, Mahwish Kundi, "Transformation of Flat File into Data Warehouse", Global Journal of Computer Science and Technology, Volume 11 Issue 13 Version 1.0 August 2011.

[2] Yang Xiang, Philip R.O. Payne, and Kun Huan" Transactional Database Transformation and Its Application in Prioritizing Human Disease Genes", IEEE, VOL. 9, NO. 1, JANUARY/FEBRUARY 2012.

[3] Alkis Simitsis, Panos Vassiliadis, and Timos Sellis "State-Space Optimization of ETL Workflows", IEEE, VOL. 17, NO. 10, OCTOBER 2005.

[4] KWANGIL KO, THOMAS G. ROBERTAZZI, "Signature Search Time Evaluation in Flat File Databases", IEEE, VOL. 44, NO. 2 APRIL 2008.

[5] Stephan Philippi and Jacob Kohler, "Automated Structure Extraction and XML Conversion of Life Science Database Flat Files", IEEE, VOL. 10, NO. 4, OCTOBER 2006.

[6] Surajit Chaudhuri, Member, IEEE, Zhiyuan Chen, Member, IEEE, Kyuseok Shim, Member, IEEE, and Yuqing Wu, Member, IEEE, "Storing XML (with XSD) in SQL Databases: Interplay of Logical and Physical Designs", IEEE, VOL. 17, NO. 12, DECEMBER 2005.

[7] Cay S. Horstmann, Gary Cornell, "Core java Volume I Fundamentals", Eighth Edition, 2008, Pages 715 – 808, 551 - 612, 361 – 492, 281 – 322, 323 – 257.

[8] Cay S. Horstmann, Gary Comell, "Core java Volume II Advanced Features", Eighth Edition, 2008, Pages 98-171, 217 – 291, 347 – 490, 588 – 600, 613 – 624, 688 – 762.

[9] Thomas Connolly, and Carlolyn Begg, "Database systems, A Practical Approch to Design, implementation and Management", Third Edition, Pearson Education, 2003.

[10] R. Elmasri, S.B. Navathe, "Fundamentals of Database Systems", Fifth Edition, Pearson Education, 2006.

[11] Abraham Silberschatz, Henry F. Korth, S. Sudharshan, "Database System Concepts", Fifth Edition, Tata McGraw Hill, 2006.

[12] C.J.Date, A.Kannan, S.Swamynathan, "An Introduction to Database Systems", Eighth Edition, Pearson Education, 2006.

[13] Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", Tata McGraw – Hill Edition, Tenth Reprint 2007.

[14] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, Elsevier, 2007.

[15] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, "Introduction To Data Mining", Person Education, 2007.

[16] K.P. Soman, Shyam Diwakar and V. Ajay ", Insight into Data mining Theory and Practice", Easter Economy Edition, Prentice Hall of India, 2006.

[17] G. K. Gupta, "Introduction to Data Mining with Case Studies", Easter Economy Edition, Prentice Hall of India, 2006.

[18] Daniel T.Larose, "Data Mining Methods and Models", Wile-Interscience, 2006.