Data Processing Model of Bank Credit Evaluation System

Guorong Xiao
Department of Computer Science and Technology
GuangDong University of Finance, Guangzhou, China
newducky@126.com

***Abstract*—Data extraction and transform are based on the mapping relationship between the new system database and old system database. In generally, transform also includes data cleaning, which is to clear the dirty data that is from original database. Before data cleaning, analysis of data quality is demanded.Data loading is to load the data which has been extracted and transformed into the destination database through loading tools and SQL sentences programmed manually. This paper detailedly analysis the model of data processing in the credit evaluation system , in which the process and methods of data extraction,transform and loading have been studied deeply, and the steps of data integration also have been presented. It also constructs a XML data model for data integration, which can implement structured data of relational type as well as describe unstructured data and semi-structured data. The model has been apllied to the bank credit evaluation system,and proved to be more efficient implementation, it can be used as the principle guiding of data processing in bank system.**

***Index Terms*—credit evaluation system, data processing model, data warehouse**

I. INTRODUCTION

The entire data warehouse system is composed of three parts: data integration, data warehouses and data marts, multidimensional data analysis. The information system which business intelligence operations rely on is a traditional system, and mutiple data sources, database and applications, the various parts can not communicate with each other. The challenge faced by enterprise application program is how to achieve correct data through different information platforms, how to integrate a large number of available data and transform them into information assets as soon as possible.These information assets allow enterprises to make more pointed responses to the market and customer needs, so as to keep a leading market competitiveness and constantly open new commercial opportunities.

Because of the source data and the format is different, so data integration is difficult. And all these require a powerful data integration technology. But as the early data sources mainly considering various relational databases, therefore, the integration is mainly pointed to relational database. For example, the integration of ODBC with JDBC is a typical integration method on relational database. Along with the rapid development of information technology, the data storage goes beyond the scope of relational database, and spontaneously the corresponding data storage technology brings a requirement of cross-platform integration of multiple types of data. And the data integration technology is still in further development.

For the enterprise,the application system currently running have spent them a lot of energy and money, in particular in the system data collection.So the purpose of the new business intelligence system is to make correct decision through data analysis. At this point, enterprises like to have a comprehensive solution to solve the difficulties and business problems of data consistency and integration, so that enterprise could collect data from a traditional environment platform and use a single solution for data conversion efficiently.

Data extraction, transform and loading solution contains three areas, first, 'selected': read out raw data from various business systems ,it is a prerequisite to all the work. Followed by is 'conversion': to obtain data conversion pumping according to the pre-designed rules, and the heterogeneous data formats here can be unified. Finally, the 'load': converting the data as planned incremental or all and import the data into the data warehouse.

Data extraction, transform is based on the mapping of relationship between the old and new system database, and data analysis is to establish mapping relations, which also includes the code-data analysis. Transform steps generally includes the process called data cleaning, data cleaning is mainly directed against the source database for the occurrence of ambiguity, duplication, incomplete, in violation of business rules or logic corresponding data. Data quality analysis is required before cleaning operation,it will identify problems in the raw data. Data loading process could use loading tool or load the data by the SQL program and then loaded the results of the data extraction, transform into the target database.

These data integration process ensure that new business data can enter the data warehouse, it blocked complex business logic of data warehouse and it provides a unified data interface, which is the most important meaning of building a data warehouse. analysis and application of users can also reflect the latest business developments. These process can be as simple as transferring data from one table to another on the same system,and it can also be as complex as taking data from an entirely system which is thousands of miles away and rearranging and reformatting it to fit with a very different system.

doi:10.4304/jsw.6.7.1241-1247

Currently, the biggest challenge on data integration process is facing with the data heterogeneity and data of low quality. A very important issue in the process of building data warehouse applications is to ensure good data quality, if the original data itself is "dirty" , it may lead to the results not match with the actual situation.

II. SOLUTION FOR CREDIT EVALUATION SYSTEM

In recent years, bank consumer credit business is booming, how to use scientific and technological means to finish personal credit rating and credit risk analysis, forecasting and assessment, are increasingly showing its necessity and urgency. Currently, the credit card system which support the daily operation of production systems and transaction-oriented daily teller operations can not provide many analysis and decision-making. Large amounts of historical data burst at the same time, and it leads to complex decision analysis. So data warehouse technology was used in the personal credit assessment and risk management of consumer credit applications. Business decisions can improve the accuracy of banks, reduce business risk, and promote the rapid development of banking consumer credit business.

A bank's total retail loans records have reached 35 million, of which 99% have been incorporated into the Consumer Credit System (CCS system) to manage.With the continuous development of the retail credit and the retail credit species, how to select target customers, target markets, how to identify, prevent and control credit risk, have become a pressing demand for credit operations. These require banks establishing personal credit evaluation system as soon as possible to adapt the existing retail credit business development. Bank credit risk management and personal credit rating system are constructed based on the consumer credit system, it anlysis the assessment of customer credit and credit quality,and construct the personal credit scoring and credit risk management systems.

To complete the customer's credit assessment and analysis of credit quality and the establishment of a personal credit assessment and credit risk management data warehouse, you need to finish the evaluation by two ways: access to risk assessment and establish monitoring and evaluation process.

For the customers who have new loans, the bank's deposits and loans system need the materials and information of these new customers. If necessary, combined with market information, it should give users a comprehensive credit score.For customers already in bank loans system or old bank credit card users, it should monitor the dynamic changes in the repayment ability of customers in order to grasp the user's credit rating degree and loan risk index.

Because of multiple business data sources and distribution of the bank's data in each sub-system, how to ensure data consistency, how to truly understand the business meaning of data across multi-platform and multi-system integration, the maximum possible to improve the quality of data, should be solved in order to meet the needs of the changing of personal credit business.

III. DATA PROCESSING MODEL

Simultaneously, as the rapid development of informatization, each unit has developed a large number of diverse application systems of software and hardware platform. And they accumulated rich data resources under a variety of application systems as well, which formed heterogeneous data because of diverse software and hardware platform and diverse data model. With the acceleration of computer's networking trend and improvement of network performance, enterprise units urgently need to integrated these heterogeneous data which are distribution in geography, autonomy in management and heterogeneousness in model, establish a unified access interface on data source logical layer, and realize the distributed sharing of heterogeneous data. To make full use of existing data resources from each information system, there often need to implement the exchange visits to data between different information systems.

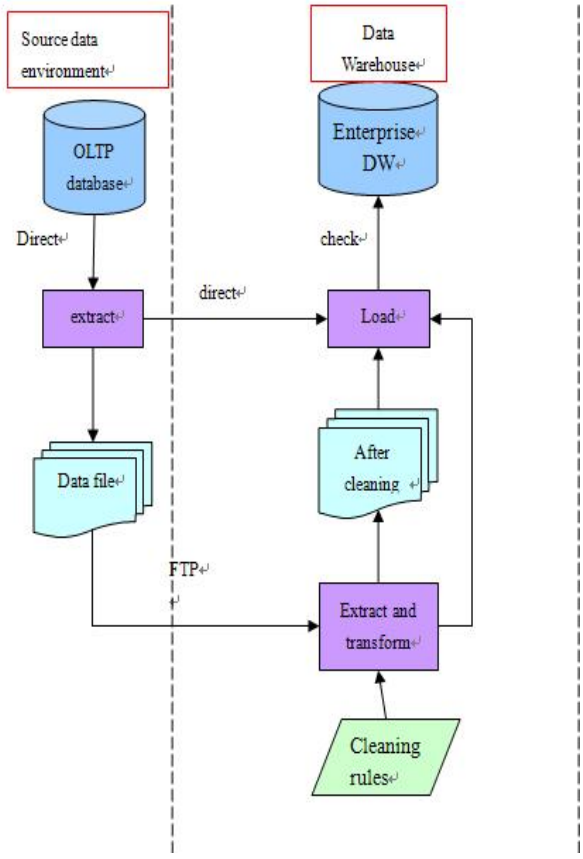

Figure 1. The data processing model for bank system

Thus the system should obtain the data in Consumer Credit System (CCS),and establish some connection with new generation of retail system (RBS) and new generation of credit card system (NCDS) , extract the required data from these two systems. RBS have a huge amount of data, one is the account and log files, if all into the data warehouse, there are more than 50G of data to download each day. If you want to finish establishing a

complete credit evaluation data warehouse with one step, in terms of resources, measured, in terms of investment and time, the difficulty is very great. Therefore, the target of the system should based on the existing business system, resolve the outstanding issues without advocating the establishment of a large and comprehensive system.So we could establish a system model as figure1 to meet the requirement of bank's data integration.

Data extraction, transform and loading architecture in the design of the bank credit evaluation system should have the following functions: management simple; using meta data method, centralized management; interface, data format, transmission are strict norms; try to install software not in external data source; data extraction system processes automation, and automatic scheduling; extracted data timely, accurate and complete; can provide the interface with various data systems, system adaptability; provide software framework for the system, functional changes, the application can adapt to new requirements with very little to change; scalability.

A. Data souce analysis

To establish a perfect credit scoring and credit risk management, apart from obtaining the data from CCS, but also from RBS and NCDS. It should establish a definite link, extract the required data from the systems.

The main source of raw data:

1) New generation of consumer credit systems: The current generation of CCS system has been basically preserved a variety of consumer credit data, which can provide some basic data source to establish a personal credit assessment and credit management of the data warehouse.

2) Bank credit scoring system: 38,000 data samples of their personal credit scores as the initial data sources.

3) Some business data entered manually: It's major role is to establish the individual customer credit scoring models and the weight adjustment.

B. Estimate the amount of data

CCS system includes consumer credit data, estimates the size of the amount of data for a total of 90G.

C. Data extraction design

Data extraction, transform and loading model is responsible for business applications in different systems of data extraction, cleansing, integration, and stored in the database.

At this stage mainly involves the following two tasks:definition of data loading, data maintenance policies. Of which:

1) Data extraction: Extract data from business applications;

2) Data cleaning: Clear the data inconsistencies in different data, verification of data;

3) Data conversion: Convert data structures and data types;

4) Data aggregation: Detailed data of the business application system needs to be aggregated into summary data;

5) Data loading: Loading the data into the database structure of data warehouse.

Data extraction is the entrance of data into the library. Since the database is a separated data environment, it needs to extracting data from online transaction processing systems (business applications), external data source, offline data storage medium into data warehouse. In general, it does not require data in the database and online transaction processing systems to maintain real-time synchronization, data extraction can be carried out regularly, but the extraction operation is performed over time, success or failure of the validity of the information in the database is vital importance. However, if database and online transaction processing system need real-time synchronization data, you can use data synchronization tool (such as MQ Series or Data replicator) to achieve this goal.

New generation of consumer credit data system is the key to get the original source data, CCS system data that is stored in the ES/9000 source data on the DB2 for VSE database.

Data collection methods are two kinds:

1) Through DB2 connect: We could use DB2 Connect to connect with the VSE DRDA DB2. This way connect DRDA host database through DB2 Connect, and then export the data of host to the RS/6000 through the export feature.This does not require programming in the host system, and with data extraction simple, and it takes small amount of work. The adoption of this approach should not affect the normal business of the host stable production as a precondition.

2) Ftp download: Extract data from the consumer credit systems by programming, generate the formation of VSAM files and download by ftp tools.This approach convert the database table file of the host system into a VSAM file by programming, then transfer the data to warehouse machine via ftp way, generate the formation of txt files. In this way the workload of the host side is relatively large, and will be impacted by the host version and need related maintenance.

In data extraction, it should consider the following:

1) The existence of manual data: The number of manual data, the existence of unstructured data, etc., extracted design work can be done only after the information data have been collected.

2) The database and storage systems handling have the same data source: This type of data source is easy for design. Under normal circumstances, DBMS (SQLServer, Oracle) will provide the database link feature,we should write select statement to establish a direct link between the relationship of database server and the original business system.

3) The database and storage systems handling have the different data sources approach: For this type of data source, in general, can also establish a database link through ODBC. If you can not create database link, you can complete in two ways, one is derived source data into .txt or .xls file through tools and then import these source system file into the ODS. Another method is to be completed through the programming interface.

4) For the file type of data source (such as. txt,. xls): We can use database tools to import the data to the specified database.Or the use of tools can also be achieved, such as SQLServer the SSIS service and other components.

5) The incremental update problem: System for the large amount of data must be considered incremental extraction. The business system will record the time which the business take place, we can use to do incremental signs. Before the start of each extract,we should determine the maximum time recorded in ODS, and then compare the time with business systems,and take all records which time is greater than it. Business systems could use the time stamp to finish incremental update.

D. Data transform

Data transform use the model of data warehouse, through a series of transformation to achieve converting the business model into data analysis model, through the built-in library functions, custom scripts or other extensions, enabling a variety of complex transformations, and supports debugging environment, a clear monitoring the status of data conversion. Data transform is the real goal of the source data into key data, which includes the transfer of data format, for data type conversion, data summary calculations, data registration and so on. However, these work can be handled in a different process, as the case may be, for example when the conversion in the data extraction, and can be converted in data loading time. Transform steps generally includes the process called data cleaning.

E. Data model for XML

For a heterogeneous data integration system, what it needs to face is variety of data sources, each of which has its own characteristics. Except for difference in data model, some of the data sources don't have a fixed model but easily vary in structure. In addition, some of them even contain a number of unstructured data. Apart from a few properties of their own, these data are difficult to be described with data model in detail.

The actual data in the bank environment, the problem most often occurs in the customer information, and customer information is critical information concerned about the banking business.Particularly,some of the original core banking system has no central customer information file, or surrounded by many independent applications, as well as bank customers have no way to easily access the omplete picture (Customer Profile).

XML is a common language specification established by W3C organization on Feb.1998, it is a simplified subset of SGML, especially designed for Web application program. It is a platform-independent representation of data. XML data can be created by any application on any platform to read. You can even manually edit and create XML code documents. XML as an extensible markup language, its self-describing property makes it great apply to data exchange between different applications, and this exchange is not based on the premise that predefines a set of data structure. The biggest advantage of XML is its ability of data description and data transmission, so it has strong openness. The tag (markup) is the key part. You could create content,and tag it with the restricted tag, so that each word or phrase could become classified information. When reading from a printout or processing documents in electronic form, the elements can help to better understand the document. It will be more easy to identify the various parts of the document if elements describe stronger,

To make it possible to exchange business data based on XML, we must realize the XML data's access in database, integrate XML data with application program and then make it combination with existing business rules. XML provides standard formats of describing different types of data-such as: appointment records, purchase orders, database records, graphics, sound and etc, it can concertedly and correctly decode, manage and display information. XML is constructed on the Unicode at the beginning, providing multilingual support with universality.

Based on these advantages of data representation, we could use XML data model to represent a bank credit client information object. For example, we create credit card files for a custmer, including the customers' information of "name", "city", "company" and etc.

First, we need to create the customer object, which represented a certain customer of the bank. A complete model presentation of customer object is showed as below:

```
<CreditCardInfo>
<Customer>
<Name>Xiaodong Chen</Name>
< IDCardNo>353389352336</IDCradNo>
< CustomerAddr> Yuexiu District </CustomerAddr>
<Nationality>China</Nationality>
<City>Guangzhou</City>
<HomeTel>0203353633</HomeTel>
<CompanyTel>0203353615</CompanyTel>
<CompanyPostcode>510000</CompanyPostcode>
<HandSetNo>1333356395</HandSetNo>
<Email>newducky@21cn.com</Email>
</Customer>
<Manifest>
<Item>
<CardAccountNo>512335338935</CardAccountNo>
<BranchCode>01</BranchCode>
<CreditLevel>A</CreditLevel>
<RiskLevelCode>133</RiskLevelCode>
<Expiration>2015-01</Expiration>
<OverdraftDays>5</OverdraftDays>
<LastBalance>2300</LastBalance>
<CurrentChange>260</CurrentChange>
<OverdraftBeginDate>2009-12-25
</OverdraftBeginDate>
<OverdraftAmount>1500</OverdraftAmount>
<OverdraftInterest>0</OverdraftInterest>
<PrimaryCardNo>512335218365</PrimaryCardNo>
<DunCount>0</DunCount>
</Item>
<Item>
```

<CardAccountNo>512335158236</CardAccountNo>
<BranchCode>02</BranchCode>
<CreditLevel>A</CreditLevel>
<RiskLevelCode>133</RiskLevelCode>
<Expiration>2013-01</Expiration>
<OverdraftDays>10</OverdraftDays>
<LastBalance>2500</LastBalance>
<CurrentChange>230</CurrentChange>
<OverdraftBeginDate>2009-10-23
</OverdraftBeginDate>
<OverdraftAmount>1000</OverdraftAmount>
<OverdraftInterest>0</OverdraftInterest>
<PrimaryCardNo>512335218365</PrimaryCardNo>
<DunCount>0</DunCount>
</CreditCardInfo>
</Item>
<Item>
<CardAccountNo>51233935633</CardAccountNo>
<BranchCode>03</BranchCode>
<CreditLevel>A</CreditLevel>
<RiskLevelCode>133</RiskLevelCode>
<Expiration>2016-01</Expiration>
<OverdraftDays>1</OverdraftDays>
<LastBalance>1500</LastBalance>
<CurrentChange>360</CurrentChange>
<OverdraftBeginDate>2010-1-25
</OverdraftBeginDate>
<OverdraftAmount>1500</OverdraftAmount>
<OverdraftInterest>0</OverdraftInterest>
<PrimaryCardNo>512335218365</PrimaryCardNo>
<DunCount>0</DunCount>
</Item>
</Manifest>
</CreditCardInfo>

The above example is simple, but it is enough to illustrate how to use XML data model to represent the data in traditional relational database completely. Through the data integration,we discover that the same customers have three different types of credit card in the bank, these all belong to one customer, so they should have the same customer information.Through this data model,credit card customer information data has been unified for data Management.

In fact, XML data model not only can represent structured data, it also can be used to represent semi-structured data and some other unstructured data. While building model with traditional relational database for the latter two situations, it will face great trouble.

Using this model,We could solve the probelm of multiple data sources on banking credit card data, and the problem of inconsistent data format, so that each customer's customer files will be unified, it is a very important role for the subsequent data cleaning.

F. Data cleaning

High quality decision-making must rely on high-quality data. In order to avoid wrong conclusions, the data accuracy is critical, otherwise there will be the so-called garbage in, garbage out phenomenon. How to detect and remove the data error,it is a key to the success of data warehouse construction.

To ensure data quality, data quality management play a pivotal role, Its goal is to ensure that the data meet user's data quality requirements. Only when we meet quality requirements of data users, we could obtain useful information from these data. If the source data itself is "dirty" data, it may lead to the result of a very different with the actual situation, this is not a technical problem but the method and application problems.

To improve the quality of the data, we general need data cleaning. Data cleaning is the technology shown with the statistics, data mining or pre-defined rules for cleaning up the data problem of source database, such as the occurrence of ambiguity, duplication data, incomplete violation of business rules or logic corresponding data,it turn the dirty data into the data meet the data quality requirement.

In order to detect incorrect data and inconsistent data types,it require detailed data quality analysis, we should use the analysis program to obtain the metadata about the data characteristics and the data quality issues. According to the number of data sources and the different level of "dirty", it need appropriate data transformation and cleaning method. Model related data transformation should use the query language statements and maps transforms automatically to generate the code. After the error is eliminated, the dirty data should be replaced by clean data.

After completion of data quality analysis,we should do data cleaning.Data cleaning is mainly used for cleaning the garbage in the data,it can be divided into before cleaning, taking in washing, cleaning after extraction. We mainly use before cleaning. The conversion of the code table can be taken into before the conversion and the conversion in the extraction process. As follows:

1) For the source database tables: According to the results of the analysis of data quality, cleaning function should be establish before data extraction. The cleaning function control program can be scheduled before the unified data extraction, it can also be distributed to each selected function scheduling.

2) For the source code tables: If no change or little change in data length, we should consider conversing the source code referenced in the data table before the extraction. We need to create code conversion function before the conversion.

3) The code quite different coding rules: We should consider finishing the conversion in the extraction process. According to the results of the code difference analysis, we should adjust all the code involved in data extraction functions.

After data quality analysis, we should generate the result of relevant data quality and the report on data quality which does not meet the analysis needs, analyze the causes of poor data quality, recommendation reports of how to improve operational quality and report of business systems.

After the research and development on the bank credit evaluation system,we could summed up some data quality causes and treatment methods in bank credit

evaluation system, it can be used as the principle guiding of data cleaning in bank system:

1) Customers Information is not standardized. Some custmers do not fill or do not fill out, resulting in the data incomplete or invalid, but these information is valuable.

2) Multiple records of one customer. The customer information at different account opening have different number of records in the bank.

3) The default value does not make sense. Such as zip code is 11111, the guarantor code 00000000, account number is 33333333333 and so on.

4) ID is not unique. An object is identified by several key identity. For example: Customer A has the customer number 1000000 in the loan system, but in the deposit system, is 20000000.His gender is identified as M, F, empty in the loan system,, but in the deposit system, identified as 1,0, blank.

5) Data missing. Some valuable information do not have detailed input. Such as gender, age, zip code, telephone number, and other family members do not have detailed input.

6) Ambiguity field. Some data field is used for a variety of purposes. The same field was re-defined as many different meanings.

7) Repeat key code. As the bank of historical data usually stored in OLTP system less than 180 days, so key code may be reused. For example, the banking institutions merge, closed 010 branches, and the OLTP system data of 010 branches will be integrated into the branch office data of 020 branches; but the historical data of 010 did not merge to branch 020. A year later, branch code of 010 was reassigned to a new branch office. Therefore, when we add historical data to the data warehouse, the new branch (010) data, there may be key code duplication.

8) Value conflicts. Some field value and record are interrelated with the values of other fields.Sometimes interdependent fields appear contradictory values. For example, a customer code for the province is Guangdong, but the Zip Code is Guangxi. Therefore, if we count the amount of the customer and give the summary of the provincial level through the provincial code and zip code, it will be two different results.

9) Violation of business rules. Such as the data values do not support the business rules. For example, overdraft account balance is positive, the error rate values, and more.

10) Data integrity problems. The data should be associated but have not associated; it should not be associated data, but linked together.

Mainly all aspects of the system data for cleaning may be ambiguous, repetitive, incomplete, in violation of business rules and other issues, the record will be removed from the database, adjust accordingly based on the actual cleaning operation.

G. Loading data

Considering with the processing of data file, a file may be assigned to three documents. Abnormal data file found the records out of the rules in the cleaning process. Valid data file is the data which is not against the rules, Abnormal data files and valid data files should be the input file. Initial summary data is loaded into the data warehouse through data conversion and cleaning, that is called storage, you can load the data file directly or use database for data loading.

Data loading has two parts, one is to load legal data file into the roll back table, the summary part is not loaded; the other is to load the initial summary table into a temporary matrix, according to the different types of data, there may be secondary aggregate. When the data to a temporary summary table is finished, the temporary table data is loaded into the middle table. The summary of the process depends only on the input file and it can not be associated with the external other files.

1) Full volume of data loading: For the small amount of data and data changes at a lower frequency of data files , such as the host system parameter file, using tools to do the necessary conversion and load it into the corresponding parameter tables & fact tables directly.

2) Incremental data loading: For the big amount of data,we should generate incremental data daily,and insert/update the data into the corresponding fact tables.

The first step is to generate incremental data file,and generate the new data checksum for each record in the new text file, then compared the value with the back up data previous day. According to the differences,we should produce daily incremental data file. Specific methods are: using the CRC32 function to generate the entire records data checksum, if checksum generated by the day are different from the same record checksum values of the previous day, then the record has been changed; If the backup checksum file have no corresponding record, the record is a new record, these two kinds of data will be written daily incremental data file (Sequence File).

The second step, load the incremental data file. After data cleaning on the incremental data file, we should load the data into the appropriate database file by insert/update way .

H. Verification after the data is loaded

After the data is loaded, you need to check the data. The results of data validation is to determine whether the new system could be used, it is an important basis for the new system opening.

The quality analysis of the data after loading could check by tools, or write a targeted inspection program. Verification of data loaded is different from the quality analysis of historical data before loading. The data validation indicators after loading include five areas: integrity check, the existence of foreign key references, consistency check, meaning of the value on the same data in different locations are the same, the number of records on the corresponding old and new database are the same; special inspection of sample data, check the same sample in new database are the same in the old database and so on.

Through the establishment of bank credit evaluation data warehouse, enterprise-level customer information can be establishde by the relevant fact table. Redetermined by the customer, the data added and with

the related accounts, transactions and other associations, banks have a good chance to achieve customer relationship management, credit scoring and customer contribution analysis.These will increase their market competitiveness.

IV. CONCLUSION

The paper detailed analysis the data extraction, data transform and data loading process, make up a data processing model for the application of bank credit evaluation system,and give detailed explanation of each step in the model. It give some innovative and effective solutions for dealing with the problem of heterogeneous data integration. Especially XML data representation technique and cross-platform distributed WebService technique based on XML, these provides a stable platform for processing problems of information representation and information access from heterogeneous data integration system.Combined with the analysis of data cleaning and analysis of data quality,it provide us a specific project design methods and technology for bank data integration system. This system model and methods have been used in application of bank credit evaluation system,and the design and the implementation have a higher efficiency, it meet the requirement of the bank on the data processing part.

REFERENCES

[1] Rahm, E., Do, H.H. Data cleaning: problems and current approaches. IEEE Data Engineering Bulletin, 2000,23(4):3~13.

[2] Surajit Chaudhuri ,Umeshwar Dayal1An Overview of Data Ware2 housing and OLAP Technology [J]1SIGMOD Record ,1997 ,26(1) :652741.

[3] Widom J1Research Problems in Data Warehousing[C]1ACM CIKM'95 ,1995.

[4] H Galhardas ,et al1Declarative Data Cleaning :Language ,Model and Algorithms[C]1VLDB 2001 ,Rome Italy ,September 2001.

[5] Bright M Wetal, "A Taxonomy and Current Issues in Multidatabase Systems". IEEE Computer, 1992, 25(3): 50-59

[6] Hecht Nielsen, R."Neurocomputing", Addison Wesley, 1990, 124-133.

[7] Laks V S , Lakshmanan , Fereidoon Sadri ,et al1 SchemaSQL :A Language for nteroperability in Relational Multi2database Sys2 tems[C]1VLDB ,199612392250.

[8] Vijayshankar Raman ,Joseph M Hellerstein 1Potters Wheel : An Interactive Data Cleaning System[C]1Rome ,Italy. VLDB ,2001.

[9] Hernandez, M.A., Stolfo, S.J. Real-World data is dirty: data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 1998,2(1):9~37.

[10] Lee, M.L., Ling, T.W., Lu, H.J., et al. Cleansing data for mining and warehousing. In: Bench-Capon, T., Soda, G., Tjoa, A.M., eds.Database and Expert Systems Applications. Florence: Springer, 1999. 751~760.[11] Monge, A.E. Matching algorithm within a duplicate detection system. IEEE Data Engineering Bulletin, 2000,23(4):14~20.

[11] Milo, T., Zohar, S. Using schema matching to simplify heterogeneous data translation. In: Gupta, A., Shmueli, O., Widom, J., eds.Proceedings of the 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann, 1998. 122~133.

[12] Lee, M.L., Ling, T.W., Low, W.L. IntelliClean: a knowledge-based intelligent data cleaner. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000.

[13] [14] Caruso, F., Cochinwala , M., Ganapathy, U., et al. Telcordia's database reconciliation and data quality analysis tool. In: Abbadi,A.E., Brodie, M.L., Chakravarthy, S., et al., eds. Proceedings of the 26th International Conference on Very Large Data Bases. Cairo:Morgan Kaufmann, 2000. 615~618.

[14] Galhardas, H. Data cleaning and integration. 2001.

[15] Li guanyu, Zhang jun, Jin qiangyong, "The Research of Heterogeneous Data Integration in Information System". ICMSE/Haerbi (2001)

[16] Galhardas, H., Florescu, D., Shasha, D., et al. Declarative data cleaning: language, model and algorithms. In: Apers, P., Atzeni, P.,

[17] Ceri, S., et al, eds. Proceedings of the 27th International Conference on Very Large Data Bases. Roma: Morgan Kaufmann, 2001.371~380.

[18] Raman, V., Hellerstein, J. Potter's wheel: an interactive data cleaning system. In: Apers, P., Atzeni, P., Ceri, S., et al, eds.Proceedings of the 27th International Conference on Very Large Data Bases. Roma: Morgan Kaufmann, 2001. 381~390.

[19] Lei qiangyong, "Research of Heterogeneous Database Integration System and the prototype of its supporting tools". Maritime Affairs University of Dalian, Master's Theses of Maritime Affairs University of Dalian, 2002, 3.

[20] Liu jun, "Prototype Implementation of Distributed Intelligent Heterogeneous Data Integration Support System". Maritime Affairs University of Dalian, Master's Theses of Maritime Affairs University of Dalian, 2003, 3.

[21] Li junhuai, Zhang jing, Zhou mingquan, Geng guohua, "Research on Method of XML-based Technology of Enterprise Disparate Data Integration". Computer Engineer, 2002, 28(9): 63-74.

[22] Chris Brandin, "XML Data Management, Information Modeling with XML", May 27, 2003.

[23] R. Goldman, J. Mchugh, and J. Widom, "From Semistructured Data to XML: Migrating the Lore Data Model and Query Language". Prpceedings of the 2nd Imternational Workshop on the Web and Databases (WebDB'99), Philadelphia, Pennsylvania, June, 1999.

[24] J. Mchugh, S. Abiteboul, R. Goldman, D. Quass, and J. Widom, "Lore: A Database Management System for Semistructured Data". SIGMOD Record, September 1997, 26(3): 54-66.

[25] S. Abiteboul, D. Quass, J. Mchugh, J. Widom, J. Wiener, "The Lorel Query Language for Semistructured Data". International Journal on Digital Libraries, April 1997, 1(1): 68-88.

Guorong Xiao received the B.E. and M.E. degrees in computer science and technology from South China University of Technology, Guangzhou City, China.

He is currently a lecturer at the department of computer science and technology, GuangDong University of Finance, China. His research interests are data mining, data warehouse and financial analysis etc.He has finished several banking, securities and mobile data warehouse.