http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (9): 1839-1844, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.1839.1844

Research of Data Quality Assurance about ETL of Telecom Data Warehouse

[1]Sun Wei, [2]Wei Wei, [2]Jing Zhang, [2]Wei Wang, [2]Jinwei Zhao, [2]Junhuai Li,
[3]Peiyi Shen, [4]Xiaoyan Yin, [5]Xiangrong Xiao and [2]Jie Hu
[1]Department of Information Engineering, Shaanxi Polytechnic Institute, Shaanxi, Xian'yang, 712000, China
[2]School of Computer Science and Engineering, XI ' an University of technology, XI'an 710048, China
[3]National School of Software, Xidian University, 710071, Xi'an, Shaanxi, Peoples Republic of China
[4]Department of Computer Science and Technology, Northwest University, Xi'an 710127, China
[5]School of Information Engineering, Zhejiang Agriculture and Forest University, China

Abstract: In recent years, with the development of data warehouse and Web technology, more and more attention has been paid to multiple applications of the data in warehouse. However, data quality issue is one of the biggest obstacles to the success using of data warehouse project for many enterprises. So a data audit model is proposed and the relevant methods are studied in the study based on the characteristics of the telecommunications industry. Further, a three data layer audit method, consisted of audit mode data file level, record level and index level, is constructed during Extraction Transformation Loading (ETL) process. It can effectively improve the data quality of data warehouse.

Key words: Data warehouse, data quality, ETL, data audit

INTRODUCTION

In recent years, China's telecommunications industry has invested a lot of manpower and material resources to proceed with the data warehouse project (Inmon, 1992), it makes people aware of the importance of information systems from a new perspective and the value of historical data (Liqiang, 2005). Since then data can not only be used to retrieve, but also be used to analyze the operational status of the entire enterprise and decide business operations (Shiwei and Qiang, 2003). But at the same time, it is also found that a consistent, clear, accurate data or good accessibility and availability of a data is the basis of the data warehouse system.

Better definition of the quality of the data is the degree of "best use" of the user, with a high degree of subjectivity and relativity (Guo and Zhou, 2002). With the development of data warehouse and Web technology and multiple applications of the data making data quality issues become the focus of much attention, data quality issues becomes one of the biggest obstacles to the success of data warehouse project of many enterprise (Jarke *et al*., 1999).

For all aspects of data quality issues, including data cleaning, data integration, detection of similar records, data quality assessment, data quality audit, there are a lot of academic researches and practical applications. The data quality audit can effectively improve data quality, control process of data quality of application system, so that the data user can understand the level of data quality of data warehouse and take appropriate measures to ensure the quality of the data (Wei and Hao, 2012).

Some researchers have done a lot of work for the enhancement of the quality of the data. some audit methods of data quality were proposed from different aspects in the literature (Wand and Wang, 1996), a tool is designed for data quality analysis and browse (Dasu *et al*., 2002), but generally they did not provide a systematic method for data quality audit (Dianc *et al*., 1997). The paper proposes a data audit model based on the characteristics of the telecommunications industry, and gives the audit method.

ANALYSIS OF DATA QUALITY PROBLEMS OF DATA WAREHOUSE ETL

Data quality problems of telecommunication data warehouse are the wrong data, missing data, data duplication, as well as the different data attributes of same data in different systems and so on (Rahm and Do, 1998). The incentive for the data quality problems can be sunmarized as the data source errors, systematic errors, rules and errors (Liu, 2009), they will be analyzed one by one as following.

Corresponding Author: Sun Wei, Department of Information Engineering, Shaanxi Polytechnic Institute, Shaanxi, Xian'yang, 712000, China

Systematic error: Due to the data of source system often changing with production, extracted data is different at different time points (Wei and Zhou, 2010). So it is required that the target system must take the last extracting time point as a starting point when extracting data. If extraction point is not exactly controlled, it is easy to cause data duplication or deletion. If the system exception occurred in the data processing, and the handling mechanism is not perfect, it can easily lead to data quality problems. If ETL system crashes during the loading process and it is lack of an effective and rapid response mechanism, it will not be able to detect and deal with the problem (Wei and Hao, 2011). If not being constrained after the abnormal data storage, it can easily lead to duplication of loaded data or missing data. So to establish a sound mechanism of monitoring and troubleshooting of ETL process to ensure that data loading failures such as the abnormalities of starting process and collapse in process can be timely detected and treated.

Data source error: Telecom has many independent application systems of business supporting prior to the establishment of a data warehouse (Rahm and Do, 2000). Data entity is maintained independently, the property values of data entities which have the same meaning are inconsistent in the different systems (Wei *et al*., 2011). For example, one product, may be encoded differently in the billing system and CRM system. There may be some incomplete data in these systems of business support , for example, name items is completed in the user profile while address data is missing (Wei and Qi, 2011). In a particular business, the lack of this type of data entity may be not important for an entire business process, but once loaded into the data warehouse, it is difficult to be perfected. Error data will be delivered to the target system if it is not discovered and excluded. Additionally, if data is input manually, a slight operation error will lead to input error data being input into the system. Therefore, inconsistent data source is another reason that can not be ignored.

Rules error: Rules errors can be analyzed from three aspects such as data integration rules, business rules and statistical analysis. Firstly, if there is vulnerability of the data integration rules, could result in data mapping error and data inconsistencies (Wei *et al*., 2012a) Secondly, in the practical application the granularity or classification of the data in the system is often different. It is because different departments have different analysis requirements for the analysis of data; Finally, due to different interpretation, definitions and calculation methods of business indicators, data re-definition appears in the process of data verification . Therefore, unified indicators must be set on the data application platform and in the production system (Wei *et al*., 2012b).

DATA QUALITY AUDIT MODEL

The data quality audit should be tried to execute in the ETL session, because every little error will be hugely magnified in the subsequent processing of ETL process. Meanwhile, the processing of the data warehouse is linearly done, it is difficult to go back to the data re-processing when an error is found (Ye *et al*., 2009). So try to eliminate errors and data quality problems in the front steps (Wei *et al*., 2010a). Through analyzing we found that the data quality problem firstly comes from source system or the extraction process, that is the error before data is processed in temporary area. Secondly, it comes from errors generated when conversion is loaded, this part of error is generated in the warehouse development process (Xu and Pei, 2011). The article proposed that the audit points can be set step by step in the ETL process according to the characteristics of business process, to strictly control the quality of the output data of various aspects, based on technical indicators and business indicators. That is, before the source data is processed in temporary area, file-level audit is applied to interface data files according to database constraint rules, to detect the legality and correctness of the data; before the data is processed in temporary area, record-level audit is applied to data table file according to database constraint rules and telecommunications business rules, to detect consistency and integrity of data. When data is loaded into the data warehouse and other layers, the index level audit should be applied according to the telecommunications business rules. The model of data quality audit is shown in Fig. 1.

Main content of index level audit is to compare calculated values of the same business indexs in the layers of the data warehouse and OLAP (Zhang and Xu, 2005), to determine whether there is missing data in the conversion process; Meanwhile, compare other reference values of business indexes to the calculated value of this index in the data warehouse or OLAP (Parsaye, 1997) to determine whether there is an exception of business index.

Data quality audit based on business rules is that use telecommunications technology to analyze specific telecommunication businesses and develop appropriate business rules in the rule base when conducting data quality audit and use these rules to complete the data quality audit work. Table 1 shows the example of the violation of the telecommunications business rules. In Example 1, the disassemble time of fixed phone is earlier

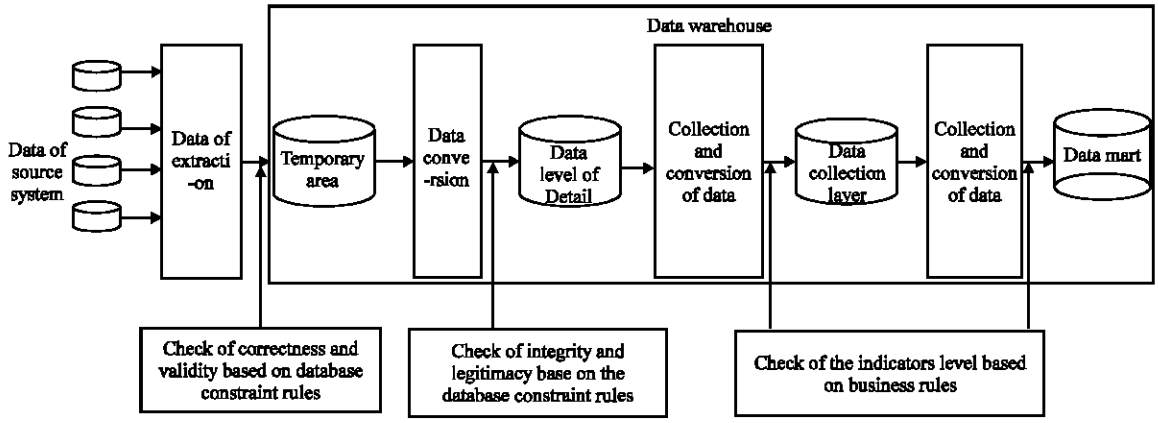

Fig. 1: Model of data quality

Table 1: Example of the violation of the telecommunications business rules

Example	Issue	Data	Reason of error
1	The phone disassemble date earlier than the completion date	Completion date =2007/02/28 disassemble date =2006/02/14	Does not meet the business rules
2	Call time is negative	Call time -20 sec	Does not meet the business rules

than installation time, where data definition is correct but it is a violation of the telecommunications business rules, so the data is wrong. In Example 2, the time value of call is negative, which is violation of the telecom business rules, so the value is a wrong data. With table, We can determine whether the value of each field in the audit record accords with business rules, for example, whether the field range of values allowed by business is beyond, can be used to determine the values of the field whether are correct or not. Data quality auditing with the business rules is also one of easiest and most effective way (Wei and Zhou, 2010).

File-level data audit: File-level data audit is shown in Fig. 2. File-level data audit is to detect and determine whether the corresponding flag file is exist, whether the flag file can be normally opened, whether control information are able to be correctly read from the flag file; wether file length value in the logo file is equal to value of total length of record of corresponding table in the interface specification multiplied by the number of records in the logo file, whether the total length of the record of the corresponding table in the interface specification is equal to total length of record of the data file; whether the size of the data file is beyond the normal range of fluctuation (Wei *et al*., 2010b).

It is also checked that whether the type, range and format match according to interface units specified in the interface specification of the recipient of data in file-level auditing. For example, it is illegal values for 202012011 as a eight bit field.

Record level audit: Record level audit is shown in Fig. 3. The main content of the record-level check is a check of the primary key; foreign key checks; coded mapping tests; data types and formats check; data value domain check; record-level business rule checking.

The record-level business rules is mainly used to audit the effectiveness of the range of the specified field of each record, record-level rules can be grouped into the single field and the multiple fields record-level rules.

The record-level rules for a single field is used to a field of a single record. For example, if the value of the i-th record "talk time" is less than 0, it indicates that the value is the error data. Its rules can be expressed as:

```
IF Reci (charge_dura) <0 THEN
The field value error data
END IF
```

The multiple fields record-level rules are used to check multiple fields of a single record. For example, the start time start_time of Ith record must be less than the end time end_time. Its rules can be expressed as:

```
IF Reci (start_time) <Reci (end_time) THEN
The field value error data
END IF
```

On record-level, whether the values of some fields in the specified table accord with the pre-configured rules in the rule table is the main work audited, the first step of the audit process is to read the constraint rules of database

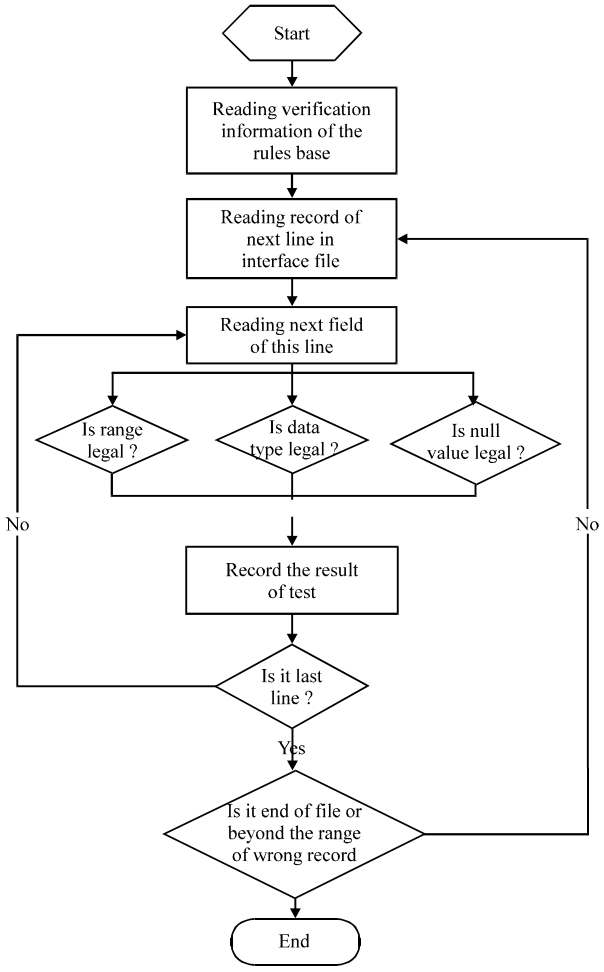

Fig. 2: Flow chart for a file-level check in

from rule base, then read the record of interface documentation, and check whether value range of every field in the input record of data files accord with business rules and whether the record length and data type accord with business rules. For example, if a field is beyond the range of value permitted by field business, you can determine whether the value of the field is correct or not. Generate test results for the record that does not meet and the file-level report for verification at last.

Index level audit: Index level checking is shown in Fig. 4. The index level audit is to audit value range and validity of index summarized and calculated on the basis of the record set, for example, compare the summary values of the same business indexes with each other in the different layers of data warehouse and OLAP, to determine whether there is omissions in the process of data conversion. At the same time, compare index references from other places

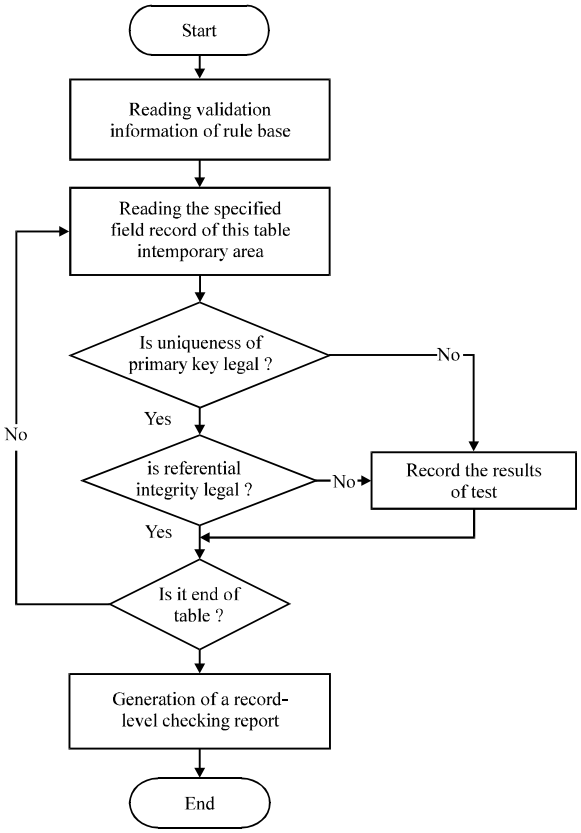

Fig. 3: Flow chart for a record-level checking

to the summary value of the index in the data warehouse or OLAP, to judge whether there is an exception in the business indexes.

The rules of index level audit can be grouped into auditing rules for single index and auditing rules for multiple indexes that are to detect business logic among multiple indexes (Wei and Zhou, 2012). Audit rules of the single index focus on the single index. For example, monthly cost of local call should not be less than 6 million according to the development of local telecommunication network. It is highly unusual if a monthly cost is less than 6 million, the reason should be found out timely. Audit rules among multiple indexes are mainly used for the detection of business logic among multiple indexes. For example, number of narrowband user in this month accords with the business logic as follows: users number of net increase = the number of new users-the number of loss of user:

```
IF (usercount! = (old_ usercount + new_user – lost_user))
THEN
The value of this field is incorrect data;
END IF;
```

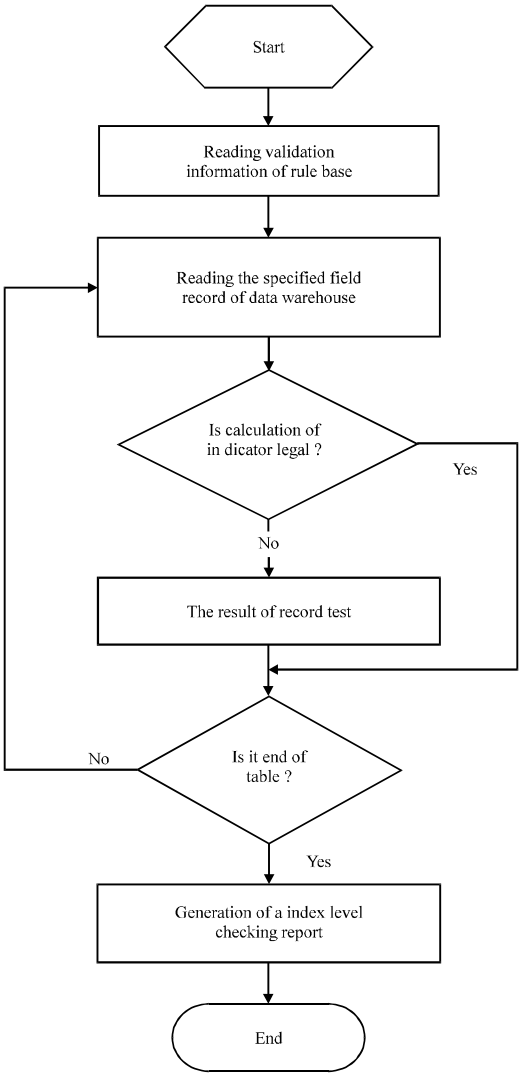

Fig. 4: Flow chart for a index level checking

Index level audit is to calculate the value of each index according to each rule in the rule table. The inspection process is to read the index level business rules from the business rules base firstly and then read the field records of the audited data table, check whether summary calculation of field records accord with the corresponding business rules, the record that does not comply with the field is recorded and the index level checking report is generated finally.

CONCLUSION

We can solve data quality problems of data warehouse not only from the source that data is generated, but also in the using link of data. But solution in using link will only get half the result with twice the effort. Only classification to perform auditing to data constantly in the ETL process can ensure the quality of the data.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable comments. This program is supported by Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No.2013JK1139). Our project is also supported by Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2012JM8047) and by Science and Technology Project of Xi'an (CX1262⑨). And this project is also partially supported by NSFC Grant (ProgramNo.61072105, 61007011, No. 61172018 81201179, 61100236, 61202393, 11226173) and Beilin District 2012 High-tech Plan, Xi'an, China.

REFERENCES

Dasu, T., T. Johnson, S. Muthukrishnan and V. Shkapenyuk, 2002. Mining database structure: Or, how to build a data quality browser. Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, June 3-6, 2002, Madison, Wisconsin, pp: 240-251.

Dianc, M.S., Y.W. Lee and R.Y. Wang, 1997. Data quality in context. Commun. ACM, 40: 103-110.

Guo, Z. and A. Zhou, 2002. Research on data quality and data cleaning: A survey. J. Software, 13: 2076-2082.

Inmon, W.H., 1992. Building the Data Warehouse. John Wiley and Scons Inc., New York.

Jarke, M., C. Quix, P. Vassiliadis and M.A. Jeusfeld, 1999. Architecture and quality in data warehouses: An extended repository approach. Inf. Syst., 24: 229-253.

Liqiang, Q., 2005. The data warehouse's construction of telecom operator in abroad. Telecommun. Sci., 1: 40-43.

Liu, Y., 2009. Data quality issues and enhance methods of telecom data warehouse. Telecommun. Sci., 9: 45-46.

Parsaye, K., 1997. OLAP and data mining: Bridging the gap. Database Prog. Design, 10: 30-37.

Rahm, E. and H.H. Do, 1998. Data cleaning: Problems and current approaches. IEEE Bull. Comput. Soc. Techn. Committee Data Eng., 41: 79-82.

Rahm, E. and H.H. Do, 2000. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23: 3-13.

Shiwei, F.Y.Y.D.T. and Z.W.Y.L.F. Qiang, 2003. Data quality managements in data warehouse. Comput. Eng. Appl., 13: 1-4.

Wand, Y. and R.Y. Wang, 1996. Data quality dimensions in ontological foundations. Commun. Acm., 39: 86-95.

Wei, W. and B. Zhou, 2010. Features detection based on a variational model in sensornets. Int. J. Digital Content Technol. Appl., 4: 115-127.

Wei, W., A. Gao, B. Zhou and Y. Mei, 2010a. Scheduling adjustment of mac protocols on cross layer for sensornets. Inform. Technol. J., 9: 1196-1201.

Wei, W., B. Zhou, A. Gao and Y. Mei, 2010b. A new approximation to information fields in sensor nets. Inform. Technol. J., 9: 1415-1420.

Wei, W. and M. Hao, 2011. Wavelet-based ARMA model application in power network. Applied Mech. Mat., 121-126: 1509-1513.

Wei, W. and Y. Qi, 2011. Information potential fields navigation in wireless Ad-Hoc sensor networks. Sensors, 11: 4794-4807.

Wei, W., H. Yang, H. Wang, R.J. Li and W. Shi, 2011. Queuing schedule for location based on wireless ad-hoc networks with d-cover algorithm. Int. J. Digital Content Technol. Appl., 5: 356-363.

Wei, W. and B. Zhou, 2012. A p-laplace equation model for image denoising. Inform. Technol. J., 11: 632-636.

Wei, W. and M. Hao, 2012. ARMA model and wavelet-based ARMA model Application. Applied Mech. Mater., 121-126: 1799-1803.

Wei, W., X.L. Yangb, B. Zhouc, J. Feng and P.Y. Shen, 2012a. Combined energy minimization for image reconstruction from few views. Math. Problems Eng., Vol. 2012 10.1155/2012/154630

Wei, W., Y. Xiao-Lin, S. Pei-Yi and B. Zhou, 2012b. Holes detection in anisotropic sensornets: Topological methods. Int. J. Distrib. Sensor Netw., Vol. 2012. 10.1155/2012/135054

Xu, J. and Y. Pei, 2011. Overview of data extraction transformation and loading. Comput. Sci., 38: 15-19.

Ye, Y., Y. Sun, B. Zhang and J. Zhao, 2009. Method research on improving data quality during data warehouse construction. Mod. Electronics Techn., 6: 100-103.

Zhang, Z. and Y. Xu, 2005. Application of OLAP technology in telecommunication field. Comput. Eng. Design, 26: 1950-1952.