A Novel Framework for Improvement of Data Quality on Integration

Zhang Guobao

***Abstract*—In the integrated situation of distributed, heterogeneous, and frequently changing data the improvement of data quality issue is more complicated and difficult. That is because there are some characteristics such as dynamic, heterogeneous, timeliness, inherent data quality etc. The issue of inherent data quality dimensions was discussed in many former literatures, but that of improvement of the integrated data quality was few. This paper proposed a framework of mastering data source, data catalog, data quality monitoring and quality visualization in order to help improve the quality of integrated data from the view of the whole systematization. The concept of data quality check point was presented and our proposed method can monitor data quality more efficiently. The functionality of framework and prototypes was presented and our approach of improvement was provided. At last our analysis was given and the proposed frame was useful and effective.**

***Index Terms*—Master data maintenance, data integration, data quality.**

I. Introduction

Data quality (DQ) referred to the condition of a set of values of qualitative or quantitative variables. There were many definitions of data quality but data was generally considered high quality if it is “fit” for intended usage in operations, decision making and planning" [1]. That meant the same data can be of good quality assessment for one usage and bad even unusable for others. This view was from usage or user perspective. The main characteristics or dimensions for DQ were: completeness, validity, accuracy, consistency, availability, and timeliness [2]. These were specific description for DQ dimensions from data self-view, inherent or external. The reasons of bad DQ problem have three main aspects: manual input or spelling for defects of information system; data input lasted a long time to incomplete data record; the ETL process led to data model changes in data integration. The dynamic changes of data included gradually input, metadata change, timeliness of data source. The mentioned problem was topical over the past few decades. Generally scholars have been solved it by two ways: syntactic control and semantic control [3].

In this manuscript the proposed approach intended creating of DQ frame model for integration of different information systems. The model was described by means of DSL (domain specific language) [4] and quality rules. The approach provided the possibility to use the frame for measurement and evaluation of data quality dimensions as well as for improving the data quality according to dynamic changes in data integration. The approach can provide visualization of the data quality result and high efficient ways. The paper deals with following issues: related conception and an overview about research in literature (Section II), a statement of problem and basic ideas for our approach (Section III), a description of the proposed frame and a design of executable check model (Section IV), and a analysis on the possibilities to adoption of our method in practice and conclusion (Section V).

II. Related Conception and Research

A. DQ Dimensions

There had been many concerned or related literature in data quality dimensions. Literature [5], [6] proposed a Multi-DimenDQ Dimensionssional data quality assessment framework, of them the most important were Timeliness, Accuracy, Completeness, Consistency, Validity.

Literature [7] argued that DQ is the multidimensional nature of the concept of quality and the requirement that data must be complied with, then ensured the quality in attendance to a specific intended use.

Literature [8] proposed a classification method for DQ dimensions, that is IntrinsicDQ, AccessibilityDQ,

ContextualDQ, RepresentationalDQ four dimensions. This was appropriateDQ description for data integration.

B. Methods for Improving DQ

To improving the quality of data and, consequently, making data fitter for use for a particular purpose. There were two main methods for improving DQ:

From the data integration view some actions aimed to improve the quality of data by preventing errors, correcting errors or proposing corrections, they were called as DQ Control [9]. This proposed approach can be implement over information system or data warehouse.

Others aimed to ensure that data selected for use have satisfactory quality for a particular purpose; this approach implied filtering and excluding data which lack the required quality for the purpose. These were known as DQ Assurance [9]. Our work belongs to the latter and mainly provides the way to make data quality assurance on integration more efficiently.

Literature [4] proposed approach provided the usage of a domain specific language (DSL) for description data quality models. Data quality models consisted of graphical diagrams, which elements contained requirements for data object’s values and procedures for data object’s analysis. This

Manuscript received May 19, 2018; revised August 17, 2018.

Zhang Guobao is with the Hohai University, China (e-mail: 20020108@hhu.edu.cn).

DOI: 10.7763/IJCTE.2018.V10.1211

approach intended to for simple data instance but not for integrated situation.

C. Data Cleaning Framework

Literature [10] defined the main four attributes of DQ and DQ was the extent that they are fitted to. It proposed a four-types classification of data quality problems according to its stem from data source or data schema. It listed server frameworks of data cleaning and argued that it was developed very difficult for data problems processing high efficiently.

III. Statement of Problem and Ideas for Approach

On integrated data situation the DQ problem became more complex than on non-integrated data situation. The DQ dimensions extended from that of centralization to distributed computing environment. More extrinsic attributes should be considered in this case. In this chapter we first generalized the intrinsic and extrinsic quality attributes and then introduced our thoughts of approach.

The analysis of these attributes can help us set up a DQ model or a global concept. Also this can give an overall understanding to the functionality of the framework.

A. Dynamics

On integration situation logically there are multiple data instances that are of central or decentralized locale distribution. The data source instance will be developed actually according to business demand and the metadata of the data source maybe change frequently or even entirely along with revolution cycle of business project. In addition the data would enter into the database gradually at the same time with business operation.

B. Heterogeneous

The multiple data instance are independent each other. They are heterogeneous systems about software product. Also there are no relevance of the metadata and business data about the two arbitrary instances although we can set up a certain relation between the selected two one.

C. Timeliness

The data has been changing continuously all the time of business lifecycle. The data inherence has attribute of timeliness and it means the time of the data to be produced or data attributes relevant to time such as operation date, birth date and so on. The every data source has various different states in according to every time snapshot like bandwidth, delay, network connectivity. The whole data source instance also has including total data amount, statistics of data item in refer to different time snapshot.

D. Inherent Data Quality

Inside database the DQ was presented as some attributes mentioned before, and mainly including timeliness, accuracy, completeness, consistency, validity, etc.

Therefore we take aforementioned statements or problems into consideration about DQ on data integration. Our proposed approach is that setting up MDS (master data system) and use it to make improvement of DQ more easily. The MDS concerns IntrinsicDQ, AccessibilityDQ, ContextualDQ, RepresentationalDQ four aspects of DQ. And we proposed concrete functionalities that are of management, of visualization, measurable. In fact the MDS provide means by which we can make improvement of DQ on data integration through DQ Control or DQ assurance methods.

The sketch of MDS was demonstrated as Fig. 1. After succeeding to make improvement by means of these corresponding methods the better quality data can be provided through data service api or interface.

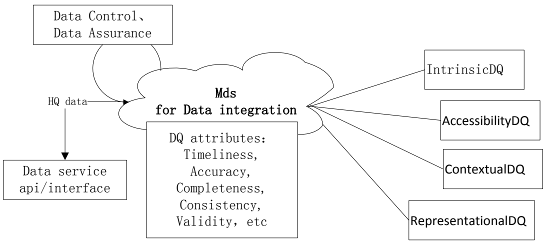


Fig. 1. The sketch of master data system.

IV. Design for Proposed Framework

A. The Functionality of MDS Framework

The framework for improvement included some functionality to achieve given goals. The functionality of MDS mainly includes that as Table I.

TABLE I: THE FUNCTIONALITY OF MDS FRAMEWORK

model	functionality	description
1	Meta-data maintenance	Register and maintenance of data source、data instance、data standard model
2	Dictionary and code standard	Maintenance of data dictionary for business data and data code of reference data standard
3	Master data management	Maintain all data records in tables that indexed in different data item or node in data warehouse. The data of every row can be retrieved in unified way according to fields of the data.
4	Data quality check	Take action to check the data DQ in reference to foremetioned dimesions of quality. That is based on rule templete can be configured to any table in different data item.
5	Data service api	Easy to build a data service api inteface presented as SOAP or JSON protocol. To provide the improved data.

The procedure of DQ improvement step by step as below:

Step 1. Register data source

Register various data source including database type like oracle, mysql, mssql etc, connection configuration information, classification description. According to this register data source connectedness monitoring can be done.

Step 2. Data integration

Use expert tool as ODI to integrate various data from different source instances.

Step 3. Master data maintenance

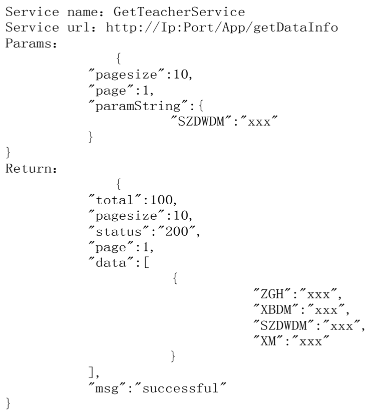
Maintain the master data from the step1 registered data source instance. The master data includes various tables or views from integrated data source. On it all data can be retrieved by unified user interface.

Step 4. Standardization and quality check

The integrated database also includes the data standard and data code dictionary. It can be compared with the master data to find out problems about data completeness and data consistency semantically. Through building data quality check model every table can be take routine check to find out problems about data completeness and data consistency syntactically. The data quality check model is illustrated as below.

Step 5. Building data service api

Based on the master data to be improved, the data service api is provided easilly by configuration step by step. The api can support JSON or SOAP protocol in order to provide web app to invoke. The created service spec demo as Fig. 2.



```
Service name: GetTeacherService
Service url: http://Ip:Port/App/getDataInfo
Params:
              {
           ″pagesize″:10,
           ″page″:1,
           ″paramString″:{
                      ″SZDWDM″:″xxx″
           }
}
Return:
              {
           ″total″:100,
           ″pagesize″:10,
           ″status″:″200″,
           ″page″:1,
           ″data″:[
                      {
                                ″ZGH″:″xxx″,
                                ″XBDM″:″xxx″,
                                ″SZDWDM″:″xxx″,
                                ″XM″:″xxx″
                      }
           ],
           ″msg″:″successful″
}
```

Fig. 2. Created service specification demo.

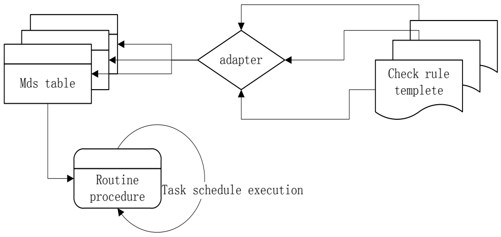


Fig. 3. The data quality check model.

Step 6. Monitoring the running service

Providing the task schedule to monitor the data service interface running and the data quality check routine execution. The result of monitoring can be sent to administrator's mailbox.

B. Data Quality Check Model

Step 4 mentioned data quality check model is shown as Fig. 3. Data quality check point was depicted as different rule template [11], [12],such as key value uniqueness, mobile telephone value, email value and so on, that can be brought out derivation from regular expression.

For example, mismatching to identity-card no rule templete is shown as:

Sql clause:SELECT * FROM @{TABLE_NAME} (1)

Where clause: WHERE not regexp_like(@{COLUMNS},"(^[1-9]\d{7}((0\d)|(1[0-2]))(([0|1|2]\d)|3[0-1])\d{3}$)|(^[1-9]\d{5}[1-9]\d{3}((0\d)|(1[0-2]))(([0|1|2]\d)|3[0-1])((\d{4})|\d{3}[Xx])$)") (2)

(1) plus (2) can form the complete rule templete, that can be matched different field of selected table through beforehand configuration operation.

The rest rule templetes can be built like identity-card no rule in the same analogy. We had these rule templetes built-in in the MDS frame.

C. The Contribution of MDS Framework

We argued that the MDS frame was helpful in improvement of data quality by several as below:

1) Concept of data quality check point

We took account the key fields or attributes of data table into potential problem points. That should be concerned than others because its importance in business. These key fields were known as data quality check point first should be mapped with quality rules.

2) Reusability of rule templete

The same rule templete can be mapped different table. This mapping can be built or deleted easily by the mouse step by step. We can also build new rule templete as regular expression format. The same rule only need to be created once and we had more time to choose which table need data quality check point according to business requirements.

3) Real time monitoring the data quality

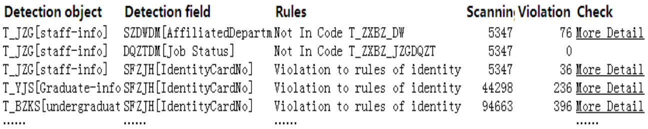
The MDS frame created several daemons to run the data quality rules configured, to run monitoring the connectedness of the distributed data sources. The daemons was set to auto run and we can run it manually by setting proper time interval.

4) Visualization of data quality

The MDS frame can show the data quality of data table in details. It included that detailed information of data source, data table, data tuple, problem data row records. This can be shown as chart or detail data list and was very helpful to identification with data quality problems. The Data quality monitoring detail list was below as Fig. 4.

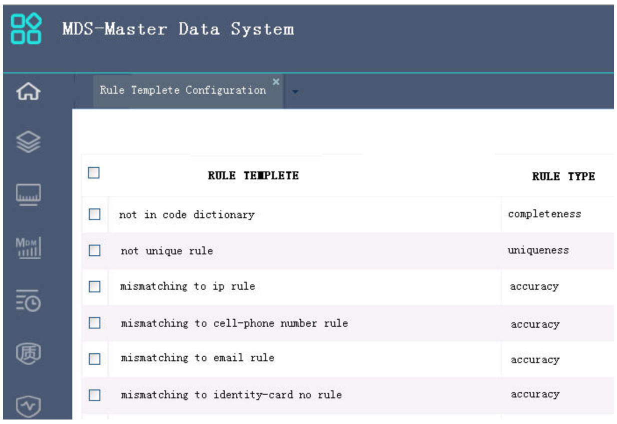
5) Generating the quality report

The result of data quality monitoring as quality report can be sent email to the administrator or other authorized user. The quality report was sent automatically or manually. The user can add or change receive mail box in the platform easily.



Detection object	Detection field	Rules	Scannin	Violation	Check
T_JZG[staff-info]	SZDWDM[AffiliatedDepartm	Not In Code T_ZXBZ_DW	5347	76	More Detail
T_JZG[staff-info]	DQZTDM[Job Status]	Not In Code T_ZXBZ_JZGDQZT	5347	0	
T_JZG[staff-info]	SFZJH[IdentityCardNo]	Violation to rules of identity	5347	36	More Detail
T_YJS[Graduate-info	SFZJH[IdentityCardNo]	Violation to rules of identity	44298	236	More Detail
T_BZKS[undergraduat	SFZJH[IdentityCardNo]	Violation to rules of identity	94663	396	More Detail
……	……	……			……

Fig. 4. Data quality monitoring detail list.



MDS-Master Data System

Rule Templete Configuration ×

□	RULE TEMPLETE	RULE TYPE
□	not in code dictionary	completeness
□	not unique rule	uniqueness
□	mismatching to ip rule	accuracy
□	mismatching to cell-phone number rule	accuracy
□	mismatching to email rule	accuracy
□	mismatching to identity-card no rule	accuracy

Fig. 5. Rule templete configuration of mds prototype.

V. Analysis and Conclusion

The proposed approach took characteristics of the integrated data condition into account. The MDS frame provides meta-data and master data register maintenance func to cover the dynamic and heterogeneous feature. That makes improvement of data quality semantically by register, classification and description. Furthermore the data quality check model solves it syntactically. The check model of templete-adapter may solve the data IntrinsicDQ problems mostly and can give the quality report explicitly and accurately by email.

The task schedule provided necessary time-sensitive routine as different running service, that includes data quality check routine, data source accessibility routine and data api monitoring routine. The timing motoring make administrator know well the data quality condition all the time along with the data integration gradually.

The data service api offered the way to simply use the improved data to a great extent. The api can be developed rapidly and adjusted refer to user's demand. The data of better quality can serve business application by this means. The prototype system of MDS frame was developed as Fig. 5. On the platform the problems of data quality can be quantized and showed in time and the users even can trace back to the corresponding data source. Our analysis illustrate the MDS frame is effective and useful in improvement of integrated data quality.

As a conclusion, the proposed framework provided the functionality of data classification management, data quality visualization, data problem traceability, data quality monitoring and data service API on data integration. The integrated data quality can be improved effectively with the help of the framework.

Acknowledgment

This work was financially supported by the research project (project number 20172067) of education information in Jiangsu, China.

References

[1] T. C. Redman, *Data Driven: Profiting from Your Most Important Business Asset*, Harvard Business Press, 2013.

[2] ISO 9001: 2015. Quality management principles. [Online]. Available: http://www.iso.org/iso/pub100080.pdf

[3] F. Boufares and A. B. Salem, "Heterogeneous data-integration and data quality: Overview of conflicts," in *Proc. 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, Sousse, 2012, pp. 867-874.

[4] J. Bicevskis, Z. Bicevska, and G. Karnitis, "Executable data quality models," *Procedia Computer Science*, vol. 104, pp. 138-145, 2017.

[5] X. O. Ding, H. Z. Wang, X. Y. Zhang, L. I. Jian-Zhong, and H. Gao, "Association relationships study of multi-dimensional data quality," *Journal of Software*, 2016.

[6] K. Yin, S. Wang, Z. Liu, Q. Yu, and B. Zhou, "Research and development on data quality assessment management system," in *Proc. International Conference on Systems and Informatics*, 2015, pp. 992-997.

[7] M. R. Bastos, J. S. C. Martini, J. R. D. Almeida, and S. Viana, *Data Integration: Quality Aspects*, IEEE, 2010, pp. 411-416.

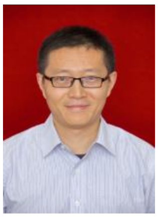
[8] G. B. Zhang and Y. J. Bian. Research on key problems of data quality control in smart campus. *Education INFO*. [Online]. Available: www.edu.cn/xxh/media/yjfz/xslt/201801/t20180129_1583024.shtml

[9] A. K. Veiga, A. M. Saraiva, A. D. Chapman, P. J. Morris, C. Gendreau, *et al.*, "A conceptual framework for quality assessment and management of biodiversity data," *Plos One*, vol. 12, no. 6, p. e0178731, 2017.

[10] Z. Guo, *et al*., "Research on data quality and data cleaning: A survey," *Journal of Software*, vol. 13, no. 11, pp. 2076-2082, 2002.

[11] J. Y. Han, L. Z. Xu, *et al*., "An overview of data quality research," *Computer Science*, vol. 35, no. 2, pp. 1-5, 2008.

[12] K. Gao, X. Diao, and J. J. Cao, "Design and application of data quality detection system based on simple rules," *Computer Technology and Development*, vol. 6, pp. 176-180, 2015.



Zhang Guobao was born in Hebei, China in 1980. He obtained the bachelor's degree from College of Computer and Information, Hohai University in 2002 and the master's degree in 2007, Nanjing, China.

His major is computer application technology. At present his research focuses on data integration, data management and mobile computing.