A Provenance-Aware Data Quality Assessment System

Hua Zheng[1,*], Kewen Wu[2], and Fei Meng[3]

[1] School of Management and Engineering, Nanjing University
[2] Department of Computer and Information Management,
GuangXi University of Finance and Economics,
Ming Xiu west road 100#, Nanning, Guang Xi, China, 530003
[3] Department of Information Management, Nanjing University,
Han Kou road 22#, Nanjing, Jiang Su, China, 210093
gxhuazheng@yahoo.com.cn, kewen-wu@163.com, fei.meng@foxmail.com

Abstract. The data quality assessment (DQA) process has the lack of sufficient attention on enterprise infomationization, and existing technologies and methods have their limitations. In order to solve data quality(DQ) problems from the source and realize the traceability of data, after research on data provenance technology and determining the idea of achieving the way data can be traced, the framework of data quality assessment based on data provenance and SOA is presented. Then the logical architecture is described, simultaneously core technology are focus to analyze. Finally, specific application is discussed and the direction of further work is given.

Keywords: Data quality management, Data quality assessment, Provenance, SOA.

1 Introduction

Data is the enterprise critical strategic resource, and reasonably, effectively using the correct data can guide business leaders make the right decisions to enhance the competitiveness of enterprises. Unreasonably using incorrect data (ie, poor DQ) can lead to the failure of decision-making. A survey of the total data quality management (TDQM) [1] project from Massachusetts Institute of Technology (MIT) is shown: only 35% of the companies trust the own data, only 15% of the companies trust the partner data. In the United States there is cost about 600 billion U.S. dollars annually to ensure DQ, or to compensate for DQ problems caused economic losses. It can be seen that the issues of DQ have been begun to attach importance by the enterprises.

Currently, most enterprise information technology projects are not starting from scratch, and they need to use the data that already resides in the enterprise and has the poor quality. In particular, because the current scope of data has been expanding and more widely shared and increasingly diverse forms of data have been, a large, complex, heterogeneous data environment has been formed. Therefore, the analysis of data

* Corresponding author.

Y. Yu, Z. Yu, and J. Zhao (Eds.): CSEEE 2011, Part II, CCIS 159, pp. 265–269, 2011.

generation and evolution process, then evaluate the quality and accuracy of data, as well as revise data result that is very important.

In this paper, firstly the background of DQA and basic ideas are introduced; In section 2, the related research works are given, which includes DQA and provenance; The framework of DQA based on provenance and SOA is designed in section 3; Provenance model is given in section 4; A concrete example is implemented in section 5; Finally a brief summary is given.

2 Related Works

Provenance is a current research focus, and how to use provenance technology to achieve DQA is a worthy research direction.

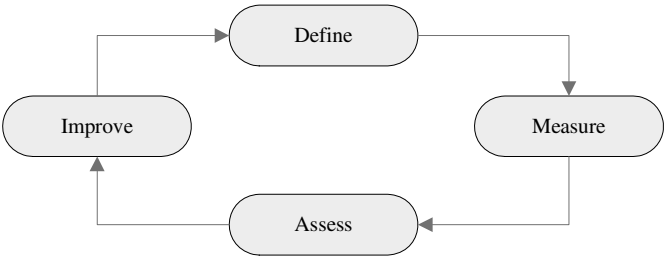

Fig. 1. Data Quality Management Process

DQ[2] is defined as "suitability for use", and this definition is now widely accepted. Much research in data quality management (DQM)[3] focuses on methodologies, tools and techniques for improving quality, and DQA [4] is an important part of DQM(shown in Fig.1). The current mainstream methods of DQA are summarized in literature[5], including: TDQM, DWQ, TIQM, AIMQ and so on, a total of 13 ways; Yang et al[6] designed a six-dimensional DQA model, and the quality situation of the data system can be assessed by the application of quantitative indicators.

Provenance[7] refers to the data are generated, and the information of the whole process of evolution over time. Data provenance contains static source data information and dynamic data evolution process, which characteristic is focused on to describe the variety application data sources and evolutionary process information, including richer metadata. After Simmhan et al[8] analyze and compare the exiting research achievements of provenance, it is given that data provenance is defined as derivative process information came from source data to data product.

3 Architecture of Provenance-Aware DQA

Realistic DQA must be carry for a specific environment and user, and there is no uniform standard. So in this paper, a framework of DQA is presented that enables the various functions of assessment to dynamically be added and in which the evolution of

the data by data provenance can be technically analyzed, and the appropriate system functionality of assessment can be selected for DQA.

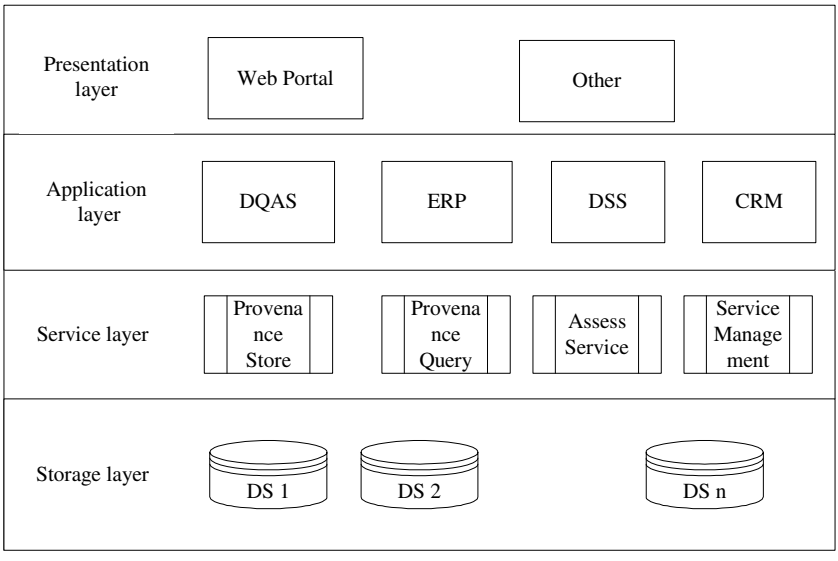

Fig. 2. Logic Architecture

SOA(Service-Oriented Architecture) is an abstract model, and represents a specific implementation which does not involve the software infrastructure and has direct service-oriented. Enterprise business logic can be achieved at lower cost for rapid reconstruction. Therefore, the assessment functions based on SOA will be abstracted in the form of service, and then the framework of DQA that can adapt to all kinds of requirement for assessment in a different environment is created. The logical architecture of the framework shown in Fig.2 is divided into four layers (storage layer, service layer, application layer, presentation layer). Its core is the service layer, in which all the functional components are packaged into the form of web service, here including the assessment of service components, provenance storage components, provenance query components, service management components. All of the services will be integrated by the ESB (Enterprise Service Bus). The entire solution is based on a typical SOA framework.

The specific process of DQA can be divided into three parts: (1)source system analysis. Data quality metrics are applied to determine data quality levels of source system; (2)target system analysis. The differences between the target system and the source system are analyzed, and then the recommendations for the elimination of differences are created; (3)alignment and harmonization requirements for each relevant data elements are assessed. The results of DQA indicators are interpreted, and translated into business terms. Detailed reports, charts and summary are created to describe DQ levels and provide recommendations.

4 Semantic Model of Provenance

In order to collect and query the provenance data in provenance management module, firstly the semantic model which can describe provenance is given. Detailed exposition is given below.

A provenance entity can be viewed as a static entity record or a activity record, which can be defined as(type, entity), type $\subseteq$ {S,A}. Suppose type=S, then the provenance entity is a static entity record; suppose type=A, then the provenance entity is a activity record. a provenance relationship is a relationship between the two provenance entities, which can be defined as(causal-entity, consequential-entity, role, annotation, relationship-id), causal-entity and consequential-entity both are provenance entity, role $\subseteq$ DICT, relationship-id $\subseteq$ NS.

There are two types of provenance entities: static entity records and activity records. A provenance entity can be accessed, so we can define the storage of the provenance entity as a physical object that can be accessed (called PES). Provenance relationship is actually the relationship between the two provenance entities, and we define a provenance relationship store here (called PRS) as a storage object of provenance relationship. Therefore, the provenance storage is actually to achieve storage of provenance entities and their relations.

An important operation on the provenance is the query, so we have developed a query model by which provenance entities and relationships can be operated. The query model is constituted by PES, PRS, and various queries operators. With comparing the general data query, the difference is that the results of a provenance query include not only the structure of complex content, but also includes the structure of complex relationship. The query of provenance information is mainly completed by defining operators of these two types of objects. TO query PES (type as a query parameter) as an example for a description: Suppose S is a object of PES, type=t, use s and t as input, the operator can be defined as:

$$\sigma_{type}(S,t) = \sigma(S, \Theta type = t) \tag{1}$$

You can use the operator to retrieve the records of all activities of PES. Because space is limited, on the other types of operator is not defined within the details.

It is known that the semantic model is constituted by provenance entities and their relationships from the above definitions. By defining the semantic model, the provenance information of all the static and dynamic elements in application system can be described by a flexible way.

5 System Implementation

Here, the DQ problems of the sugar factory are selected to study. In the infomationization process of sugar factory, a distributed control system (DCS), a equipment management system (EMS), a condition monitoring system (CMS), a enterprise resource planning (ERP) system and so on are respectively constructed. For the different business lines, channels or products categorization, the data is often stored and processed in different ways by using different technologies. Core enterprise information is distributed in multiple vertical systems with multiple copies. The

maintenance information of each system is according to their own context, regardless of the context of the whole enterprise, which further exacerbates the inconsistencies in the process of business.

Based on the solution that was discussed in the previous sections, we developed a manufacturing-enterprise-oriented DQA system for the sugar factory, in which the data of management and control will be monitored by provenance management, and the problems of DQ caused by misuse and false behavior will be found. The dimension of evaluation can be customized by user. The major algorithm of evaluation is using simple ratio, maximum/minimum operation, weighted average, etc. This system provides strong support to enterprise DQM and contributes heavily to control/ management integration.

6 Conclusions and Future Work

Because it is not feasible for universal dimensions and methods of DQA, so the dimensions should be analyzed for different application environments and individual, then the appropriate assessment methods are selected. In this paper, the proposed DQA framework based on SOA and provenance is a feasible solution, in which assessment function is dynamically adjusted by the form of web service.

The future works will be focused on further optimizing the details of this solution, including structural provenance model, automatically collecting and storing provenance data, and how to secure the provenance data, etc.

Acknowledgment

This work was supported in part by a grant from the Research Foundation of Philosophy and Social Science of GuangXi Province, China (No.08FTQ001).

References

1. Wang, R.Y.: A Product Perspective on Total Data Quality Management. Communications of the ACM 41(2), 58–63 (1998)
2. Wang, R.Y., Strong, D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4), 33–50 (1996)
3. Evena, A., Shankaranarayananb, G., Bergerc, P.D.: Evaluating a Model for Cost-Effective Data Quality Management in a Real-World CRM Setting. Decision Support Systems 50(1), 152–163 (2010)
4. Pipino, L., Lee, Y., Wang, R.Y.: Data Quality Assessment. Communications of the ACM 45(5), 211–218 (2002)
5. Batini, C., Capplello, C., Francalancl, C.: Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys 41(3), 40–52 (2009)
6. Yang, Q.Y., Zhao, P.Y., Yang, D.Q.: Research on Data Quality Assessment Methodology. Computer Engineering and Applications 40(9), 3–4 (2004) (in Chinese)
7. Buneman, P., Khanna, S., Tan, W.C.: Why and Where: a Characterization of Data Provenance. In: 17th International Conference on Data Engineering, pp. 316–330. ACM Press, London (2001)
8. Simmhan, Y., Plale, B., Gannon, D.: A Survey of Data Provenance in E-Science. ACM SIGMOD Record 34(3), 31–36 (2005)