Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Grid-based Data Quality and Data Integration Research

Xingman Li

College of Computer Science and Software, Tianjin Polytechnic University

Tianjin 300600, China

E-mail: simen_ok@163.com

Chunqing Li

College of Computer Science and Software, Tianjin Polytechnic University

Tianjin 300600, China

E-mail: frankly_lcq@163.com

Zhiyong Wang

Maton Information Technology Service (Tianjin) Co. Ltd

A303 Golden Sail Building, 18 Han Kou Xi Road, He Ping District, Tianjin 300057, China

E-mail: wangzy@tjmaton.com

Abstract

Data information is important strategic enterprise resources, rational and effective use of the correct data to guide business leaders to make the right decision-making, enhance the competitiveness of enterprises. The data quality and the data integration, very important speaking of the enterprise, is the enterprise innovation development power. In order to improve the efficiency of data integration, we must permit multiple applications to share computing resources. The grid technology may step isomerism platform computing resource to carry on the work distribution and to carry out, uses the existing hardware property effectively or new highly effective, the economical hardware, may carry on highly effective to the data, expands economically, so that adjustment and optimization face enterprise's data transmission.

Keywords: Data quality, Data integration, Data clean, Data grid, Data transmission, Isomerism platform

1. Background

In the present era, the enterprise informationization's request is getting more and more urgent, a very important aspect is the business data management. For most enterprises, ensure that data quality is a formidable challenge. PricewaterhouseCoopers issued the whole world data management survey result indicated. 75% of the companies believe that data lacking can lead to serious problems; over 50% of the companies overrun the cost due to the inner; over 33% of the companies can not but retard or give up use the new system's plan; over 20% of the companies thought that is unable to satisfy the contract or the agreement service level. [1] Up to 2009, Because of neglects the data quality question, some 50% above data warehouse project is unable to obtain the customer approval, even is defeated completely.

Although some projects might not overrun at all, good business planning requires taking this into consideration as part of the overall plan. It is a great challenge that must be faced to improve the data quality and reduce IT costs. The relation between improving data quality and data integration is interdependent. Improving data quality can make data integration more exact, whereas, we can improve data quality of a system with the help of data integration. Also, we can improve the data quality in the process of data integration. This both already may parallel, may also carry on separately (Ralph Kimball (2008)).

The main goal of gridding is to support coordinated work with the share source, which attribute to the result that the study on gridding data management has become very hot (Informatica (2008)). The effective and economic tensility for data integration can be achieved by developing the gridding computer system. Commercial hardware, for example, has shown the great demand for the tensility to reduce costs evidently. However, these griddings are dynamic for the nodes keep increasing and decreasing. Besides, collateral projects need to be timely adjusted in order to achieve the high-point, and reduce the load and frequency of modification of data mapping in order to answer the changing situation.

2. Data Quality Management

2.1 Summarization of Data Quality

In a very long time, data quality, also called intrinsic quality, mainly denotes the quality produced in the process of data production, such as precision, conformity, and integrality. The concept of data quality has been enlarged with the accumulation and widespread application. The contentment for the users' demands is the important target to scale the data quality.

In this meaning, data quality is a general term of a pair or a group of specific data precision, including the way of data input and flow in companies. Companies may not know the whole impact from the low or unknown data quality on the industrial operations if the definition of data quality is too narrow.

2.2 Standard for Measurement of Data Quality

2.2.1 Completeness

Measuring data needs elementary data elements, including correlative definitions and context-sensitive information for understanding and explaining data.

The process to create one master record may mean compiling or consolidating data from many records into single or multiple systems. Opportunities to perform consolidation or de-duplication must be considered (Keim DA, Panse C, Schneidewind J, Sips M, Hao MC, Dayal U. (2003)).

2.2.2 Consistency

There will always be challenges around alignment of data that need to be taken into account. Differences in data models across systems will lead to challenges with alignment of the structure of data and the actual data model to be used. Conformity can be used to measure whether tables in the database meet the specific regulations.

2.2.3 Conformity

Differences in data models across systems will lead to challenges with alignment of the structure of data and the actual data model to be used.

2.2.4 Integrity

It denotes the extent of information integrity, including entity integrity, citation integrity and domain integrity. Entity integrity requires each line in a table must be exclusive; citation integrity defines a cited relation between correlative rows in different tables of Relational Data System (RDS); domain integrity requests a row of data in the table needs to be in the legal data bound. The calculation of integrity is as follows: In the data set all satisfies the condition (to be possible to be above three one) in the data quantity/set records total *100%.

2.2.5 Mutuality

Most of the applications require accessing a certain data bound. To support specific applications, correctness is the essential attribute of data quality. Integrity, consistency and conformity all reflect the correctness in many aspects (Mike Schiff. Data Quality First(2006)).

Integrity tests the data correctness from the legal point of data numerical value; consistency relying on whether data accord with the application of logic; conformity from the lifecycle of the special product-data.

The relations between the characteristics of data quality are as follows: (Figure 1)

3. Data Integration Analysis

3.1 Summarization of Data Integration

Data integration is a process in which data from different sources and formats are integrated, logistical or physically. Data integration is traditionally divided into data warehouse and FDBS. Data warehouse technology centralizes data from many data sources into a central database physically. Whereas, only by interpreting users' queries into data sources queries, FDBS can integrate data logistically.

3.2 Federal Database System

FDBS is made of half-self-ruling database. Due to screening the differences from all kinds of the data sources, it can take a real-time and shortcut manipulation on the data of isomerous data sources (IQ Insight (2006)). Meanwhile, in FDBS, all the data sources provide the interfaces for the mutual accesses. FDBS can be centralized database, distributed database or other federal database.

3.3 Data Warehouse

An authority in the data warehouse filed called W.H.Inmon gave a short but comprehensive definition: data warehouse is a thematic, compositive and non-losable data aggregate, a decision-making process supporting management

department (Guo QJ, Yu HB and Wu K (2005)). In the process of enterprise management and decision-making, it is a thematic, compositive, time-relating and non-modificated data aggregate, in which data is classified as a generalized, function-independent and non-overlapped theme.

3.4 Middleware model

The above-mentioned methods, to some degree, solve problems the data-share and intercommunicative aspects, but at the same time, it also exists the following differences: FDBS mainly faces the integration among many databases, in which data sources may be mapped to every data mode. The enormous compositive system will endanger big difficulties in actual developing (Porto F, Silva VFV, Dutra ML, Schulze B.(2005)).

4. Informatica Data Quality and Data Integration Typical Solving Scenario Analysis

4.1 A Show of Scenario Flow Chart

Enterprises need to know more about the data in the source system. It is necessary for the enterprises to be able to integrate the data from many systems into a newer and more effective data intensity application procedure, as well as cleanse and enhance data.

To support today's business processes and goals, all corporate data needs to be universally accessible, flexible, reusable, and certifiably accurate (Mike Schiff. Data Quality First (2006)). Organizations need to know more about what is in their source systems, It is necessary for the enterprises to be able to integrate the data from multiple systems into new, more productive data-intensive applications, and they need to be able to cleanse and enhance data as well as monitor and manage the quality of data as it is used in different applications.

The platform of Informatica Data Integration and Data Quality provides the function of data analyzing, exploration, cleansing, conversion and coordinating, which can cooperate with each other in the different periods of data quality and data integration flow (Informatica (2008)). On this condition, it can provide the right and enterprise data quality service in the unification compositive environment. The concrete flow chart is as follow:(Figure 2)

4.2 FlowChartAnalysis

Data quality starts from the understanding of all the data in the system. According to the flow chart, we can see that accessing data, in batch and real-time modes, becomes especially important with the increasing data. Once the problems of data quality are found, it can be given a timely validation and correction to cleanse data, whereafter the well-cleansed data of high quality is transformed and reconciled to be integrated into the data system (Keim DA, Panse C, Schneidewind J, Sips M, Hao MC, Dayal U. (2003)). By doing that, it can make sure the data consistency as well as generating data monitoring report.

5. Conclusion and Outlook

There is no systemic appraisal target of data quality at present. The present data quality evaluation only aims at the important qualitative index, such as the problems of consistency, integrity and complexity. Data quality has become the increasingly serious problems which both big and small corporations are facing. It is vital to all the proposal of data integrity. The well data quality can furnish the corporations with supports on the decisions, becoming a new impetus for the corporations' innovation. Reversely, the cheesy data quality can bring unexpected resistance and even disaster. Improving data quality is a very long period pursuit; meanwhile, it needs incessant efforts and experiences. We should believe that data quality and data integrity will gain further development with maturity the Internet and cloudy calculation technology.

References

Guo, QJ, Yu, HB and Wu, K. (2005). "Research & application of distributed condition-based maintenance open system". Computer Integrated Manufacturing Systems, P416-421 (in Chinese with English abstract).

Hailong, Ting, and Hongbing, Xu. (2007). *Data quality analysis and application;* Computer technology and development NO.3 VOL.17;P1-3 Workshop, P 45-57.

http://www.businessobjects.com/pdf/products/eim /iq_insight.pdf.

Informatica (2008). "Making Data Work: Addressing Data Quality at the Enterprise Level"; Informatica White Paper; P2-4.

IQ Insight. (2006). "Data Quality Assessment Solution [EB/OL]".

Keim DA, Panse C, Schneidewind J, Sips M, Hao MC and Dayal U. (2003). *Pushing the limit in visual data exploration*: Techniques and applications. In: Günter A, et al., eds. P 37-51.

Mike Schiff. Data Quality First(2006)."It's Just Logical[R]".Business Objects White Paper.

Porto F, Silva VFV, Dutra ML and Schulze B.(2005). *An adaptive distributed query processing grid service.* In:

Pierson JM, ed. Data Management in Grids: 1st VLDB

Ralph Kimball. (2008). “An Architecture for data quality”. A Kimball Group White Paper.

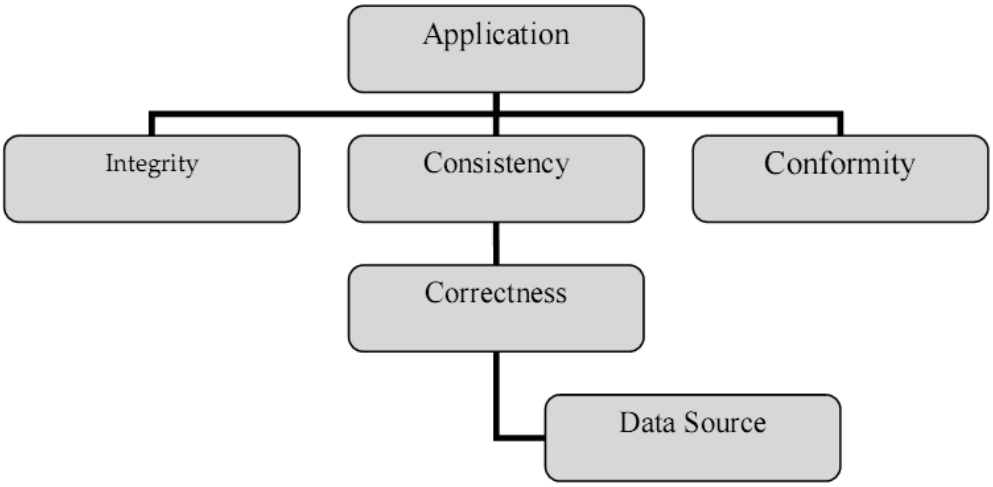

Figure 1. the characteristics of data quality chart

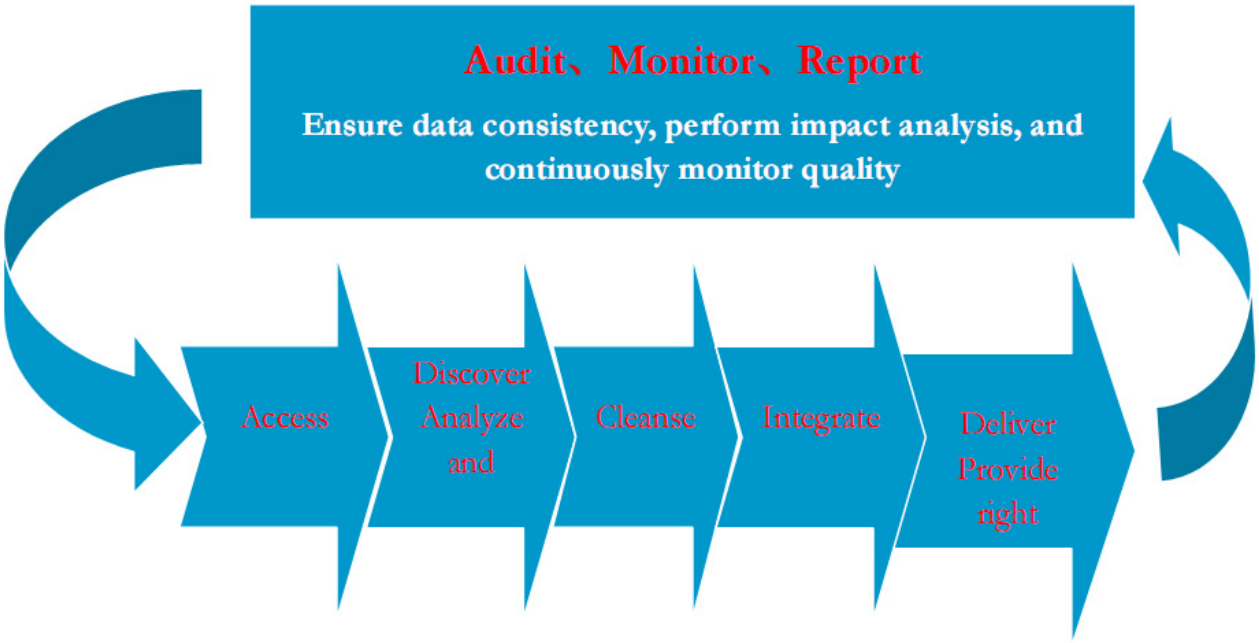

Figure 2. Scenario Flow Chart