Intelligent Information Management, 2019, 11, 24-31
http://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

Application of Integrated Reorganization of Science Specimen Data Using Kettle

Zhiyuan Wu[1], Yang Mei[2*]

[1]Information Network Center, China University of Geosciences, Beijing, China
[2]Institute of Earth Sciences, China University of Geosciences, Beijing, China
Email: zywu@cugb.edu.cn, *yangmei@cugb.edu.cn

How to cite this paper: Wu, Z.Y. and Mei, Y. (2019) Application of Integrated Reorganization of Science Specimen Data Using Kettle. *Intelligent Information Management*, **11**, 24-31.
https://doi.org/10.4236/iim.2019.111002

Received: October 26, 2018
Accepted: January 1, 2019
Published: January 4, 2019

Abstract

Standards and specifications are the premise of integrated reorganization of science specimen data, and data integration is the core of the reorganization. ETL [1] which is the abbreviation of extract, transform, and load [2], is very suitable for data integration. Kettle is a kind of ETL software. In this paper, it has been introduced into the integrated reorganization of science specimen data. Multi-source and heterogeneous specimen data are integrated using kettle, and good results have been achieved. It proved the effectiveness of kettle in the integrated reorganization of science specimen data. The application has practical significance, and the method can be referenced when reorganizing other resource data.

Keywords

ETL, Kettle, Science Specimen, Integrated Reorganization, Data Integration

1. Introduction

For many years, many units of China have collected and collated a large number of science specimens. With the advent of the era of Big Data, digitization and informatization of science specimens are advancing fast, and the volumes of specimen data are growing rapidly. Although China has developed data specifications for eight major areas of specimen resources, how to implement efficient integration of scientific specimen data through technical means and achieve long-term and continuous integration is a technical difficulty. In the past, data integration usually utilized the functions of the database itself, such as triggers, PL/SQL stored procedures, DBLINK and other functions to complete the extraction, query and association of the required data. The integration process is basically manually programmed by software developer, and the maintenance cost is

higher in the later stage. When the requirements of data integration change, the developer needs to modify the source code of the program. It is difficult to guarantee the efficiency when the requirements of data integration become complex and the amount of data become increasingly large. ETL which is the main technology to build data warehouse and realize data integration can solve many problems faced by the integration of scientific specimen data. It can realize efficient and continuous integration of massive and heterogeneous specimen data, and has advantages in applicability and performance. In this paper, kettle, a kind of ETL software, was selected to effectively integrate specimen information resources, and standardized integration of scientific specimen data has been realized. It is of great significance for the establishment of resource database for scientific specimen, and is of great benefit to the information sharing and effective utilization of specimen resources.

2. Kettle

2.1. ETL Technology

ETL represents the technology implementation process from data extraction to loading. It is a data integration tool often used in the integration of heterogeneous multiple data sources. ETL technology can transform data from data source to target data warehouse [3]. The transformation is an important step to implement resource integration. There are usually three processes in ETL. The first is extraction which is the premise of all work. In the process, data is read from decentralized heterogeneous data sources. The second is transformation which means transforming the extracted data to uniform standard result data following the pre-designed transform rules. The third is loading. In the loading process, the transformed data is loaded into the data warehouse incrementally or fully as planned. The processing flow of ETL is that: first, a verification step is used to determine what type of data is reached or extracted, and then the data is sent to a specific transformation to be processed. When the transformation is completed, the data is passed to the next transformation or to a target table. In the case of an error, it is transferred to an error process for processing. The concept of ETL is shown as **Figure 1**.

Study on the process of ETL focuses on the research of data extraction and data transformation. For complex data, Strong *et al.* [4] defined a common data quality, which can make ETL work efficiently. Kimball *et al.* [5] put forward 6 important indicators of data quality evaluation in ETL. Wang *et al.* [6] propose an ETL service framework based on metadata, and the metadata is used to describe the structure of the data warehouse and the establishment of the method.

2.2. Kettle

At present, software companies have developed many outstanding ETL systems, such as Oracle's Oracle Warehouse Builder (OWB) [7], Microsoft's SQL Server Integration Services (SSIS) [8], IBM' Data Stage. These commercial softwares

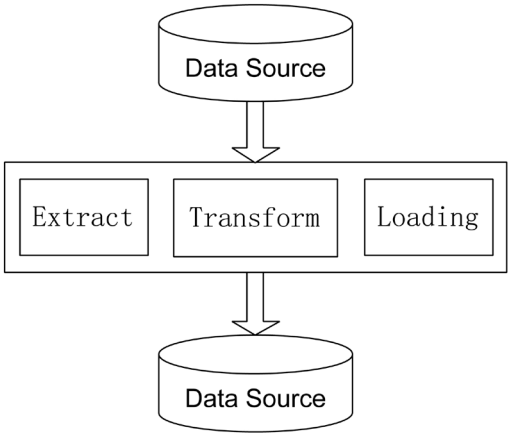

Figure 1. Structure diagram of ETL concept.

have their own advantages and the application scope, but they are expensive. Kettle [9] is an open source ETL software and it is free. It is written in pure java, and can run on the platform of Window, Linux and Unix. It is easy to use and is widely used in data transformation. The function of data extraction of kettle is efficient and stable. Graphical user environment is provided to describe what you want to do, not how you want to do it [10]. Kettle allows you to manage data from different data interfaces, and supports most input and output formats, such as text files, data sheets, free or commercial database engines, etc.

3. Reorganization Scheme of Science Specimen Data

The key to the reorganization of science specimen data is to integrate data of the same elements from multisource of different spaces, time. Data reorganization needs to focus on the determination of the specimen data element object, the complete set of the element attribute first, and then unify the semantic standard, range scope and numerical unit of the attribute item. To implement the integration of science specimen data, standards and specifications must be first. Data standards and specifications that have been formulated by state authorities can be referred to in Reorganization. The Reorganization scheme of science specimen mainly includes database design, software tools, implementation of integration.

3.1. Database Design

Database design generally has six stages: requirements analysis, conceptual model design, logical model design, physical design, test modification, and data dictionary writing. The design of science specimen database follows these six stages, mainly based on the existing technical standards of science specimen data. The process comes with analyzing the element objects and attribute fields, and designing a standardized element data table structure.

3.2. Software Tool

Technical methods that are usually used for data reorganization include writing

code by hand, and using tools software. When writing code manually, it is easy to make mistakes. Because of lack of consistent logging and error-handling, code maintenance is difficult, and performance problems may occur when integrating data with large volume. Kettle has good performances in data source support, data transformation, data management and scheduling, integration and openness, and metadata management, etc. Using kettle as an ETL software tool can reduce the effort of writing code and can satisfy the requirements of data multi-source, heterogeneous, integration efficiency and performance in data reorganization.

3.3. Implementation of Integration

Through analyzing the data source, elements and attributes, content format of the original science specimen data, etc., the data transformation workflow can be designed for various data sources in kettle according to the technical standards and database design. By running kettle jobs, operations such as data extraction, merging and integration are performing. The science specimen data will be integrated into the target database with unified standards. The target database will provide basic data sources for data management, data sharing, data mining, and data analysis of science specimen resources.

4. Example: Reorganize Mineral, Rock and Fossil Specimen Data

Mineral, rock and fossil specimen resources are classified as one of the eight major resources in the field of natural science and technology resources in China [11]. They belong to scientific specimens. This chapter will take mineral, rock and fossil specimen data [12] as an example to illustrate the process of reorganization of scientific specimens.

Mineral, rock and fossil specimen data comes from many different resource units, so distributed processing is adopted to implement reorganization. That requires resource units to send the specimen data to centralized specimen data center or provide data interfaces, and then the data center uniformly reorganizes the source data.

The flow of mineral, rock and fossil specimen data reorganization is designed as follows: first, design database according to the standard specification related to rock mineral fossil specimen resources; then design data integration process in kettle; finally, execute the transformation or job and implement data reorganization in kettle. The application of kettle to implement data reorganization is the emphasis of the whole flow, and it is also the focus of this paper. Reorganization process of mineral fossil rock and specimen resource data in kettle is shown in **Figure 2**.

4.1. Target Database Design

Data dictionary of mineral, rock and fossil specimen resource database must be

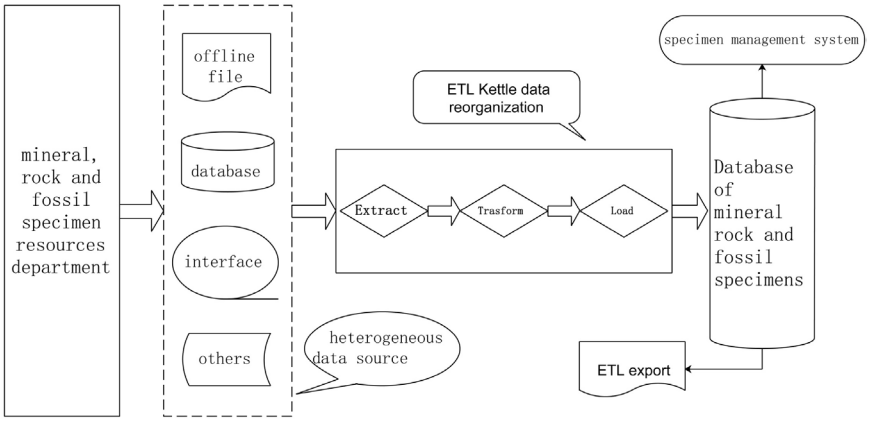

Figure 2. Reorganization process of mineral, rock and fossil specimen data in kettle.

designed prior to using kettle. Data table refers to the "mineral rock and fossil resources description specification" is the core table of database. The data of the specimen database are grouped by tabular data and image data. They are as follows:

1) Tabular data: Data are stored in two-dimensional tables, including specimen data sheets, specimen hierarchical classification and coding table, and resource unit table. The mineral fossil specimen data table is the core table of the database, which has 29 fields. The classification and coding table refers to the classification specification, and its structure is organized as the form of directory tree.

2) Image data: Image data are files which are pictures of the mineral rock fossil specimen taken by camera. File name of the picture is saved in the image field of the specimen data table.

4.2. Data Source Preparation

Since the specimen resources come from different units, the data is discrete and heterogeneous, and the degree of association is low. Data format, content, method of collection, and data quality of the source data must be understood before data integration. The way in which the resource owner units submit data can be summarized into the following ways.

1) Offline file mode. The submitted data is offline file which is usually Excel, Access, Txt and other data file.

2) Database connection mode. Relation database such as oracle, MySQL, SQL server, etc. is used. This mode connects database using database connection string, and has security risks because of exposing the database connection parameters to public. It is suggested to be used temporarily or internally.

3) Online interface mode. Interface services are provided by the resource units through http protocol. Current popular interfaces are Rest and Soap. Interfaces are easy to use relatively, but the technical requirement for developing interface is high.

4.3. Kettle Integration

Kettle has good support for the data sources above. It has multiple input components is built in which support different types of data sources, shield the differences between different data sources. Thus, unified data view for input can be acquired.

Kettle has a graphical user interface called spoon which is the design environment for creating transformations and jobs. Transformations and jobs can be quickly designed in spoon, and saved in the kettle repository for later reuse. The design of the integration procedure in kettle is as follows:

1) Data extraction. It is necessary to select the corresponding built-in input components according to every data source in kettle, and design separate integration process respectively. Full or incremental extraction can be selected as needed. The way that full amount first and increment later is generally chosen.

2) Data transformation. Data transformation requires establishing data mapping first. Specimen data of some units are strictly following the reorganization standard. They can be transformed by establishing data mapping. However, there are many specimen data whose data formats are different from the reorganization standards. These data usually come from existing database or files, and they do not conform to the standard specifications. According to the data, transformations must be designed elaborately to unify the data standard.

For example, some specimen resource units provided data source which are database connection mode or online rest interface. The data transformation designed in Kettle is shown in Figure 3.

3) Data Loading. After data have been extracted and transformed in kettle, they are stored into the target database designed before. The target database follows the reorganization technical standards fully and can be loaded when needed. So data reorganization from multi-source heterogeneous format to a unified database is realized. The target database provides a standard data source for resource management system, and can be exported to other data format such as Excel file.

4.4. Results of Reorganization

According to the established reorganization scheme, data source of specimen

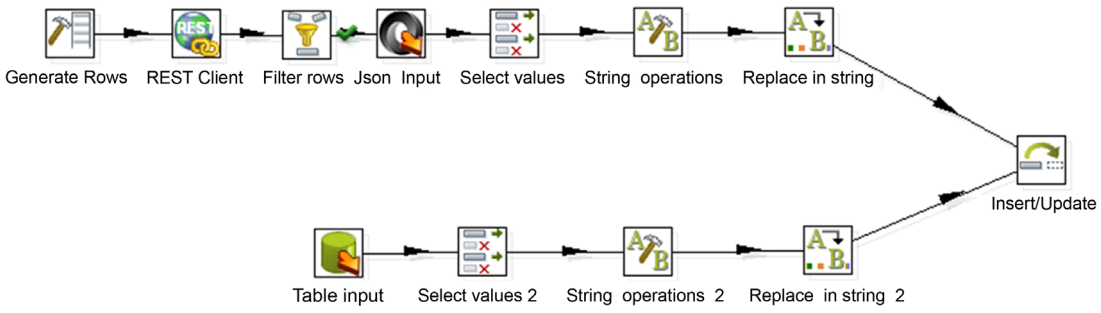

Figure 3. Data transformation flow.

resource is analyzed first, then data integration flow is designed in kettle. After the execution of transformation, kettle was successfully used to realize the standardized integration and reorganization of specimen data. After a lot of tests, it shows that if the amount of input data is greater, the integration efficiency is higher. For example, the operation time of kettle is generally less than 10 seconds if the amount of input data is approximately equal to 1000, and is generally no more than 30 seconds when the amount of input data is approximately equal to 10,000. Compared to manual programming, integration with kettle is very fast.

Saving the Kettle integration procedure in the resource repository has some benefits. The designed transformation can be reused directly if the subsequent data source is not changed in format. If change occurs, the procedure can be adjusted according to the change of data source. Because there is no need to redesign the process, integration efficiency will be improved.

During the integration process, some factors may cause an exception in the execution of the job. The integrity and consistency of the data need to be checked during the operation of the kettle job. If the quality of the data source is poor, abnormal termination of the operation may occur. The exception will seriously affect the efficiency of integration. Through analyzing the exception log of kettle, it can be decided that which part of the source data resulted in the problem. Problems will be fed back to the resource owner units, which will require resource units to modify the problem data.

5. Summary

For the integration of science specimen data, because of the multi-source heterogeneity, the work of the reorganization becomes complex. Kettle can efficiently extract, transform and load science specimen data from heterogeneous data sources. It satisfies the requirements of data integration in standardized integration and reorganization.

In this paper, mineral, rock and fossil specimen data reorganization was taken as an example for the reorganization of science specimen data. Kettle was used to design and realize the data reorganization with heterogeneous specimen data sources. The result shows that reorganization by using kettle can solve the problem that data is difficult to achieve integration efficiently when the specimen data are multi-source and heterogeneous.

Because of the multi-source heterogeneity of science specimen data and the complexity of ETL technology, further application and research need to be conducted to innovate more common method for data integration in kettle, which will make reorganization more automatic, efficient and intelligent.

Acknowledgements

It is a project supported by the National Infrastructure for Science and Technology—National Infrastructure of Mineral, Rock and Fossil Specimen Resources.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Miao, J.J., Deng, S. and Liu, Q.B. (2004) Overview on ETL Technology. *Computer Engineering*, **30**, 4-6.

[2] Xu, J.G. and Pei, Y. (2011) Overview of Data Extraction, Transformation and Loading. *Computer Science*, **38**, 15-20.

[3] Xu, H.H., Lian, L. and Yao, H.L. (2018) Establishment of Meteorological Data Warehouse Based on ETL Tools. *Computer Systems & Applications*, **27**, 224-228.

[4] Wang, R. and Strong, D. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information System*, **12**, 5-33. https://doi.org/10.1080/07421222.1996.11518099

[5] Kimball, R. and Caserta, J. (2007) The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. *Journal of Information Management*, **23**, 1123.

[6] Wang, H.M. and Ye, Z.W. (2010) An ETL Services Framework Based on Metadata 2*nd International Workshop on Intelligent Systems and Applications*, Wuhan, 22-23 May 2010, 1-4. https://doi.org/10.1109/IWISA.2010.5473575

[7] Borowski, E. (2008) Design of a Workflow System to Improve Data Quality Using Oracle Warehouse Builder. *Journal of Applied Quantitative Methods*, **3**, 198-206.

[8] Haselden, K. and Baker, B. (2007) Microsoft SQL Server 2005 Integration Services. E'earson Education, Beijing.

[9] Casters, M, Bouman, R and Dongen, J.V. (2014) Pentaho Kettle. Publishing House of Electronics Industry, Beijing.

[10] Yin, X.N., Zou, X.T. and Zhang, D. (2013) Kettle-Based Data Extraction and Transformation during Water Sector Census in Beijing. *China Water Resources*, **21**, 57-59.

[11] Wang, Y.H., Zhang, W. and Shen, X.Y. (2008) Research and Practice on National Infrastructure of Natural Resources for Science and Technology of China. *China Science & Technology Resources Review*, **40**, 16-19.

[12] He, M.Y., Yang, M. and Wu, Z.Y. (2017) Construction of National Mineral Rock and Fossil Specimen Resource Sharing Infrastructure. *e-Science Technology & Application*, **8**, 24-31.