CHEMICAL ENGINEERING TRANSACTIONS

VOL. 51, 2016

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Tichun Wang, Hongyang Zhang, Lei Tian
Copyright © 2016, AIDIC Servizi S.r.l.,
ISBN 978-88-95608-43-3; **ISSN** 2283-9216

DOI: 10.3303/CET1651076

A Design of ETL for the Construction of Traffic Network Based on Big Data

Qinan Liu

College of energy and transportantion engineering，Jiangsu Vocational Institute of Architectural Technology, Xuzhou,
Jiangsu 221006, China
lqnch@163.com

Intelligent traffic management system based on big data is the trend of the development of the transportation system. It integrates information technology, wireless communication technology, computer technology, sensor technology and other advanced technologies to form a comprehensive and efficient integrated traffic management system. Traffic system based on big data needs to deal with a large number of unstructured data and semi-structured data. Also, the capacity of the data becomes larger, the data grows faster, and the format of the data becomes more complex. Traditional ETL technology has been unable to meet the needs of the construction of intelligent traffic network which is based on big data. According to the characteristics of the traffic network based on big data, this paper designs a kind of ETL system with high universality and high data processing efficiency. First, in order to improve the efficiency of data processing, we optimize the workflow of ETL. In order to make ETL suitable for big data traffic network environment, we redesign the ETL data processing rules by identifying and merging. And then we optimize the extracting, transforming and loading of the ETL system. Finally, the experimental results show that the redesigned ETL system can effectively serve the traffic network system based on big data. This method has a high efficiency in processing complex data structure and large data capacity of big data.

1. Introduction

With the rapid development of the city, traffic congestion, traffic pollution and traffic accidents are the major problems to be solved in the city. Intelligent transportation system based on big data is imperative. It embeds the internet of things and car networking technology into traffic management, so it can control the entire traffic management system efficiently and comprehensively. However, due to the rapid increase of traffic users, the traffic related data of intelligent traffic system has jumped from the TB level to PB, EB level, even ZB level. A great amount of data has a great impact on the operation and management of intelligent transportation system. Traffic data includes data base, picture, video, text and other structured data, unstructured data and semi-structured data and so on. Researchers in the field of big data generally believe that data mining and analysis is the key of big data technology. However, many scholars in the study find that the implementation time of the data acquisition and preprocessing stage usually accounts for 60%-80% of big data entire process time. This stage is the most time consuming stage in the whole system development process (Abiteboul et al., 1999). In the process of data acquisition and preprocessing, we need to extract data from distributed and heterogeneous data sources, and then clean, transform and integrate the data. Finally, we load the data into the data warehouse. This process is called extraction, transformation and loading (ETL). Due to the large number of data from different sources, the work performance of intelligent traffic system based on big data is directly influenced by processing efficiency and quality of ETL.

The traditional traffic control system mainly uses the ring coil and video to detect the traffic flow, but the system is passive. The whole framework of intelligent transportation based on big data includes physical sensing layer, data processing and analysis prediction, and optimization management application. According to the wide application of big data systems in variety of fields, the traffic network architecture based on big data is shown in Figure 1.

Please cite this article as: Liu Q.A., 2016, A design of elt for the construction of traffic network based on big data, Chemical Engineering Transactions, 51, 451-456 DOI:10.3303/CET1651076

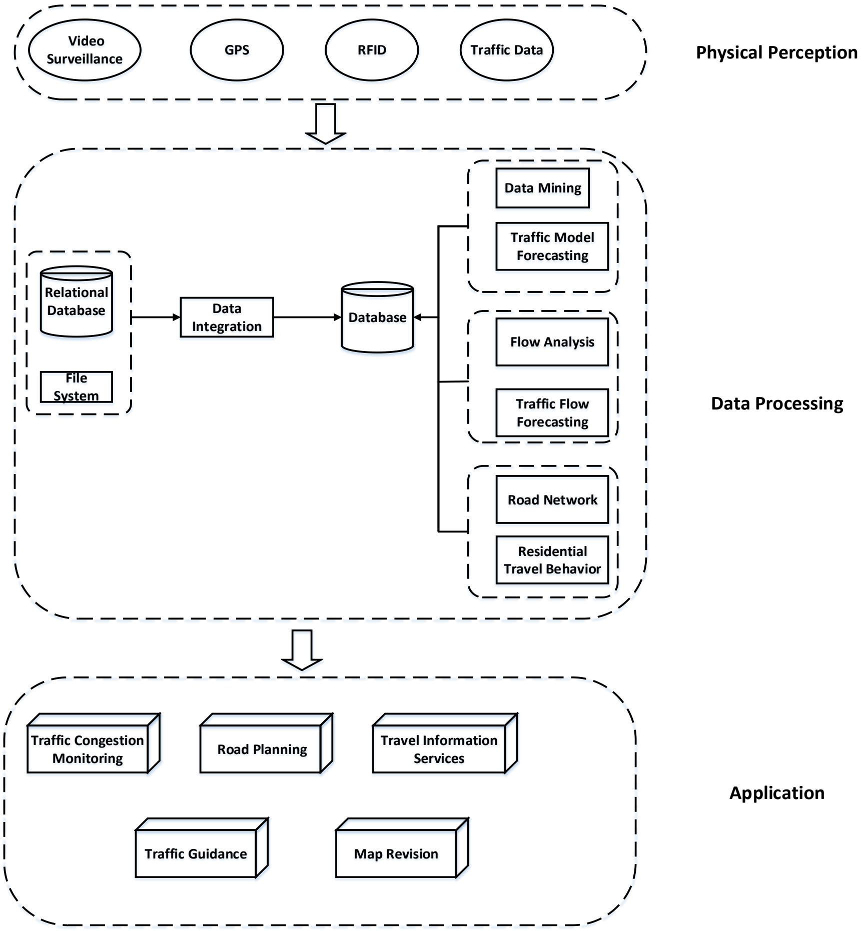

Figure 1: The traffic network architecture based on big data

The big data technology can be widely used and widely recognized in traffic network, what is mainly reflected in the real-time distribution, efficiency and predictability of big data technology. The real-time and distributed data processing guarantees the high efficiency and predictive ability of the big data in the traffic network application. Excellent ETL design is the premise to guarantee the real time processing of the distributed traffic data. At present, the major software companies have developed many outstanding ETL system, such as Oracle's Oracle Warehouse Builder (OWB) (Borowski et al., 2008), Microsoft's SQL Server Integration Services (SSIS) (Haselden et al., 2007), IBM' Data Stage. There are single ETL tools, such as Kettle, Clover ETL, Talend and Data Cleaner. These soft wares have their own advantages and the application scope. Researchers in the study of ETL mainly concentrate in 4 areas: ETL modeling, ETL processing, data quality and metadata (Xu et al., 2011). Modeling research mainly focuses on conceptual modeling, logical modeling, model transformation and optimization etc. Mu et al. (2009) propose a method for automatically generating code in the process of ETL, which is based on conceptual model. Lenzerini et al. (2003) study the definition, initial loading scenarios and conceptual model specification, and improved the concept of modeling. Ding (2007) simplifies the general model of ETL activities, and models the workflow on basis of the process. Study on the process of ETL focuses on the research of data extraction and data transformation. Zhang Rui (2010) describes the core contents of ETL from extraction, transformation and loading of three different angles. Zhang Xufeng et al. (2006) divide data extraction into total data extraction and incremental data extraction, and the ETL process is divided into full ETL process and incremental ETL process. In the face of complex data, Strong et al. (1996) defined a common data quality, which can make ETL work efficiently. Kimball et al. (2007)) put forward 6 important indicators of data quality evaluation in ETL. Wang et al. (2010) propose an ETL service framework based on metadata, the metadata in data warehouse is used to describe the structure of the data warehouse and the establishment of the method.

2. ETL technology

ETL is an important part of the construction of data warehouse (Zhang Ning et al. (2002)). Firstly, it solves the problem of data dispersion (Ma (2004)). Secondly, ETL can not only filter the "dirty data" in the data set, but also can guarantee the data clean and effective (Lin (2003)). ETL is the process of data extraction, data transformation and data loading, every industry will encounter the problem of mass data processing. Figure 2 shows the structure of the traditional ETL model.

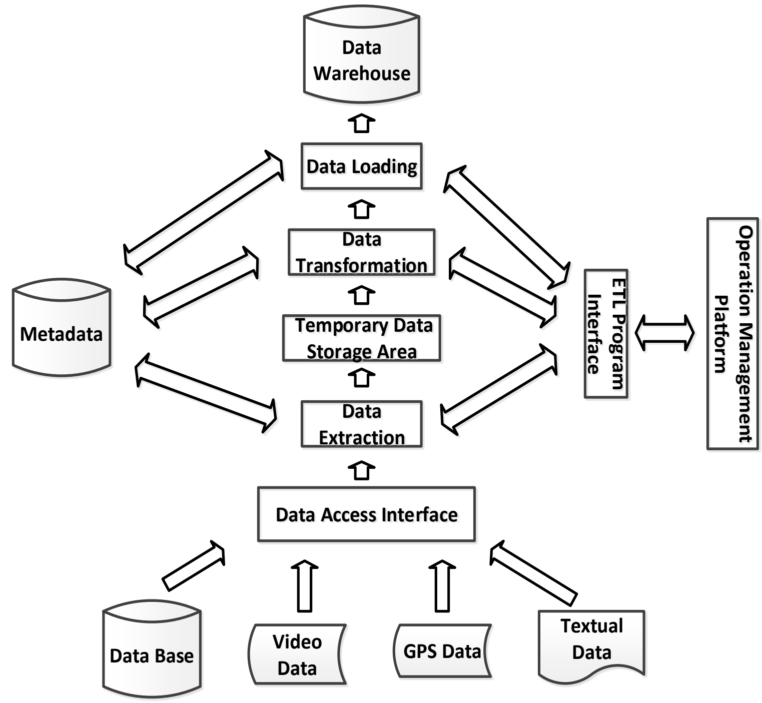

Figure 2: The structure of the traditional ETL model.

2.1 Data extraction

Data extraction is the process of extracting data from the source database system. The data that is extracted from different databases or heterogeneous sources is the basis for subsequent data processing. Data extraction has two modes, namely incremental extraction and total extraction. Incremental extraction is the most important way to extract data.

2.2 Data transformation

The result of the data transformation is that the data required by the database is obtained after the transformed. Data transformation focuses on the data processing, and it contains a lot of data processing rules. Common rules for handling data include data extraction rule, data loading rule, filter rule, the projection rule, column splitting rule, rule to merge column, column function rule, column filtering rule, table joining rule, multi broadcast rule, search and replace rule, derived column rule, column mapping rules, order rule, and group rule.

2.3 Data loading

Data loading refers to loading transformed and processed data in the activity area into the user specified database. The whole process of ETL is over until the completion of the procedure.

3. Traffic network data analysis

The construction of intelligent transportation network makes the traffic data more and more diversified. The continuous video signal of camera, the text information of traffic police portable equipment, and traffic illegal captured pictures produce a large number of types of data. These data are not only complex but also have huge data capacity. The core of ETL technology is loading data from the source system into data warehouse in accordance with certain rules and logic, which is used to further analysis of data mining.

4. The design and implementation of ETL for the construction of traffic network based on big data

4.1 Optimal design of ETL

Workflow is a summary model of workflow execution process. In this paper, a reasonable set of ETL data processing rules can achieve the optimization of workflow and improve the efficiency of data processing. In addition, we add a temporary use of the database in the middle of the database, namely the memory database, so that the frequent operation of the data can be completed quickly in the memory.

1) Optimizing the setting of data processing rules. Firstly, we classify the rules in accordance with their handled data size into different types. These rules acting on the same data type are the first class, such as column splitting rule, column function rule, column filtering rule, and search and replace rule etc. The rules acting on the single data set are the second class, such as filter rule, the projection rule, multi broadcast rule, column mapping rules, group rule etc. Third type of rule can act on two data sets, namely, the join rule. Then, in this paper, these three categories of rules are classified, the first type of rule is denoted as CRule rule, the second type is Output rule, the third type is Group rule, and the fourth type is Join rule.
2) In order to ensure that the data transformation can quickly extract the required data in memory, instead of the disk database I/O operation, we put all kinds of data temporarily stored inside the memory. This method is convenient for data transformation, and greatly reduces the time cost for disk database I/O operation. This is the purpose of adding temporary database. The optimized ETL design architecture is shown in Figure 3.

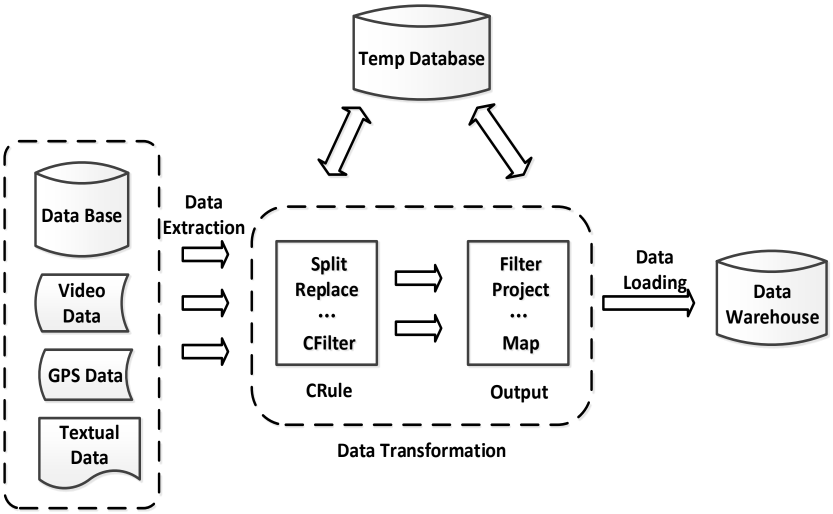

Figure 3: The optimized ETL design architecture.

4.2 Optimal design of ETL

According to the redesign and optimization of ETL structure, this paper implements a strong data processing capability, real-time and efficient ETL tools. ETL first extracts different types of data sources. Before the execution of the workflow, the process is first analyzed. Then, according to the type of conversion, the different conversion is sent to different execution parts. ETL workflow execution starts from the Begin node and ends at the End node.

5. Experiment and analysis

5.1 Introduction of experimental environment

This experiment compares the improved ETL with the traditional ETL, the results of experiments shows that the method has a significant improvement in the processing of big data. In order to verify the performance of the improved ETL, the configuration of the computer we choose is as follows: Intel(R) Core(TM) i5-4300M @2.60GHz; RAM: 4G; OS: windows 7; jdk7.0. Database using My SQL 5.0, SQL Server 2008, Oracle 10i. Data comes from a city near the second ring traffic network information in 2013.

5.2 Experimental results and analysis

1) Firstly, we compare the efficiency of ETL and traditional ETL in data conversion. The experiments are divided into 5 groups. Each group is divided into 5 experiments, which is based on the number of rules. The numbers of these rules are 2, 4, 6, 8, and 12. The test of each rule is carried out for 2 times, and the average time of the test is taken. As can be seen from Figure 4, the execution time of ETL in this paper is obviously less than that of the traditional ETL tools.

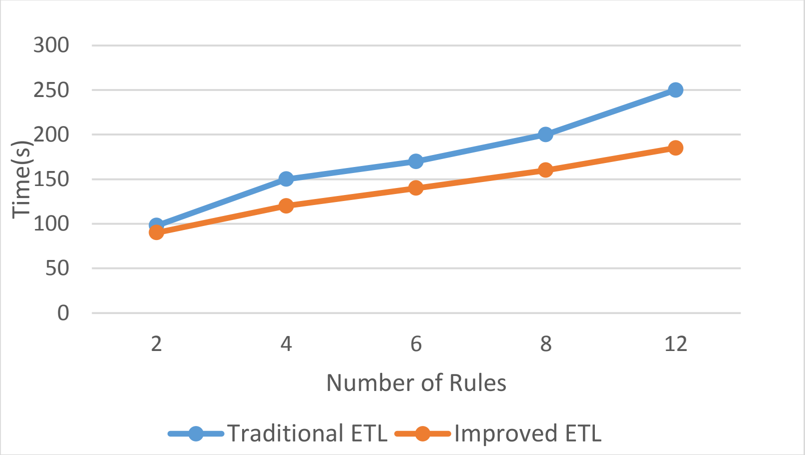

Figure 4: Comparison of ETL time

2) The advantage of the improved ETL tool in this paper is that it has a fast response time when the data is processed. And at the same time, the ETL can quickly extract, transform and load the useful data into the data warehouse. In this experiment, we compare the throughput of the data with two different ETL tools, and verify the processing ability of the different ETL tools to the data. As can be seen from Figure 5, the processing capacity of the method is not obvious, even worse than the traditional method, when the data size is small. However, with the increase of data size, the advantages of this method in the data processing ability are reflected.

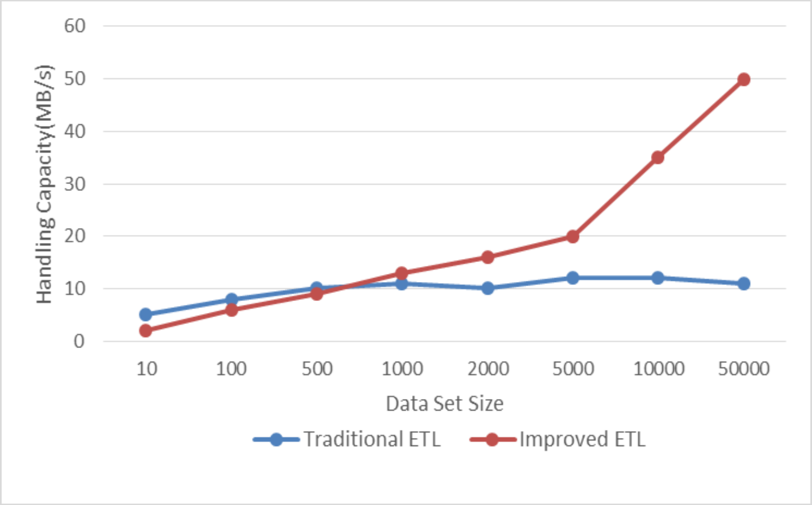

Figure 5: Comparison of experimental results

6. Conclusions

By optimizing the data processing rules of the ETL data transformation process, and adding a memory database in this process, we propose a design and implementation of ETL for the construction of traffic network based on big data. Compared with the traditional ETL tools, the ETL can process a variety of data types, and transform data faster. In this paper, the ETL mainly deals with the large amount of data generated by intelligent traffic. First of all, we prove this method to achieve the desired purpose, saving more time during data conversion. Secondly, we verify the method in the ability to handle large amounts of data processing. With the increase of data size, the advantages of this method in the data processing ability is reflected. Finally, the ETL tool designed in this paper can not only meet the requirements of data acquisition and pre-treatment of big data of traffic network, but also has very good performance.

References

Abiteboul S., Cluet S., Milo T., et al. Tools for data translation and integration. IEEE Data Engineering Bulletin, 1999, 22(1): 3-8.

Borowski E.L., 2008, Design of a workflow system to improve data quality using oracle warehouse builder [J]. Journal of Applied Quantitative Methods, 3:198-206.

Ding Z.M. 2007, Modelling and optimization of ETL workflow: [D]. Shanghai: Shanghai Normal University

Haselden K., Baker B., 2007, Microsoft SQL server 2005 integration services [M].E'earson Education India.

Kimball R., Caserta J., 2007, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Journal of Information Management. 23(11):1123

Lenzerini M., Vassiliou Y., Vassiliadis P., 2003, Fundamentals of data warehouses [M].Springer.

Lin Y. et al., 2003, Principles and practice of data warehouse [M], Beijing, Post & Telecom Press.

Ma H.L., 2004, Research on data acquisition and modelling in data warehouse [J], Journal of Minzu University of China (Natural Sciences Edition), 13(4):339-342

Mu L., Maz O., Trujillo J., 2009, Automatic generation of ETL processes from conceptual models. in: Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP. 2 Penn Plaza, Suite 701 New York NY USA: ACM. 3340

Wang H., Ye Z., 2010, An ETL Services Framework Based on Metadata in: 2nd International Workshop on Intelligent Systems and Applications. Wuhan, China: IEEE Computer Society. 1-4

Wang R., Strong D., 1996, Beyond accuracy: What data quality means to data consumers. Journal of management information systems. Journal of Management Information System. 12 (4):5-33

Xu J.G, Pei Y., 2011, Data ETL research review. Computer Science, 38(4): 15-20

Zhang N., Jia Z.Y., Shi Z. Z., 2002, Research on ETL technology in data warehouse[J], Computer Engineering and Applications, 24:213-216

Zhang R., 2010, An overview of ETL data extraction [J]. Software Guide. 9(10): 164165

Zhang X.F, Sun W.W, Wang W. et al., 2006, Research on the generation method of incremental ETL process automation [J], Journal of Computer Research and Development, 43(6): 10971103