ATLANTIS PRESS

Advances in Engineering Research, volume 170

7th International Conference on Energy and Environmental Protection (ICEEP 2018)

Research on Key Technologies of Railway Master Data Management for Big Data Applications

Yifei Liu[1, a], Ping Li[2, b], Lianbao Yang[1, c]

[1] Postgraduate Department, China Academy of Railway Sciences, Beijing 100081, China

[2] Innovation Center of Railway Big Data Research and Application, China Academy of Railway Sciences, Beijing 100081, China

[a]364270257@qq.com, [b]ict-liping@sina.com, [c]yanglianbao_121@163.com

Keywords: Railway Master Data Management; Big Data Application; Data Modeling; Data Standard

Abstract. With the advent of the big data era, the application of big data technology in the railway industry is becoming more and more widespread. Through the analysis and application of big data, the management level and economic benefits of the China Railway Corporation can be effectively improved. In order to achieve a good analysis and application effect, it is necessary to unify the basic data in each business system of the railway. As the core data that can support the basic data of railway big data applications, railway master data plays an important role in the big data application of railway. Therefore, this paper proposes the overall framework of railway master data management from the perspective of big data applications. And it studies key technologies of railway master data such as data modeling, data cleaning, maintenance procedures and version management. Finally, it introduces the application of master data standard in big data integration. And this is of great significance to the extensive application of big data technology in the railway industry.

Introduction

In recent years, information technologies such as the Internet of Things, cloud computing, big data, and artificial intelligence are triggering a new round of scientific and technological revolutions and industrial revolutions. The Party Central Committee and the State Council attach great importance to informatization, aiming to build a nation of strong network, and fully promote the development of informatization. The "Internet +" plan has fully entered the implementation phase. In the field of railway, with the continuous improvement of the application of informatization and the further completion of the infrastructure in railway information, the total amount of data resources in the railway industry has also increased substantially. However, in terms of the management, analysis, and utilization of railway data, there still exist several problems. The standard of the basic data is not uniform, the quality of original data collection is not satisfied, the intercommunity among different systems can be more close and the application of big data technology on the railway industry is insufficient[1,2]. Therefore, in order to improve the degree of railway information sharing, fully embody the value of data, and further increase the strength of data-driven business, China Railway Corporation has projected and built a Railway Data Service Platform[3]. It covers basic data management, data integration, data sharing, big data storage and analysis, and provides services for big data applications such as data sharing interaction and deep data value mining.

As the core basic data supporting the analysis of railway big data, the railway master data describes the core business entities which is needed and shared across railway departments. As an important platform for unified management of railway master data, the railway master data management platform[4] is an indispensable part of the basic data management platform in the railway data service platform. It aims to provide unified and authoritative master data for railway data service platforms. The railway master data management platform can also improve the quality of railway data and the accuracy of big data analysis. Therefore, this paper proposes an overall framework of master data management and a master data model based on knowledge graph from the perspective of big data applications, and makes a study of the key technologies.

General Framework for Railway Master Data Management of Big Data Applications

The goal of the railway data service platform is to provide the services of basic data, shared data, and big data analysis for each business application system, which is mainly composed of data integration, data sharing, big data storage and analysis, and fundamental data management. The data integration module mainly provides functions such as structured data integration, real-time streaming data integration, and unstructured data integration to meet various requirements of data collection. The data sharing module is mainly oriented to the needs of storage, query of full-type data (structured, semi-structured, and unstructured), and realizes the storage and calculation of structured data and unstructured data. It can conduct the function of management such as the application and authorization of the data tables. Then generate different data sharing strategies for different users, granularity being controlled to the field level. The module of big data storage and analysis mainly provides statistical analysis, data mining services, visualization services, and other functions. It uses data warehouses and data marts to build statistical analysis functional components for business users. The module of basic data management mainly provides the functions of master data management, geographic information management and metadata information management, which are used to meet the requirements of unified and centralized management and service of the basic data. The geographic information data management provides public geospatial data for various application systems, realizing the sharing of railway geographical information. The metadata management defines and manages data elements and relative information, including data formats, aliases, and information sources, etc.

The railway master data management takes the railway master data management platform as the carrier to connect the basic, highly shared, and low-frequency data in the railway business systems as master data to the master data management platform. It also has achieved centralized cleaning, maintenance, and management of railway master data with the help of key technologies and management process such as master data integration, data storage, and master data management[5], So the railway master data management can ensure the consistency, accuracy, integrity, and relevance of railway master data, and also provide authoritative and standardized master data. The data quality of the railway data service platform is improved by adding master data standards to the master data fields contained in the business data. The railway master data management ensures the accuracy and effectiveness of the data in the process of big data's analysis and application, which makes the results of the big data analysis more precise. The general framework for railway master data management of big data applications is given in Figure 1.

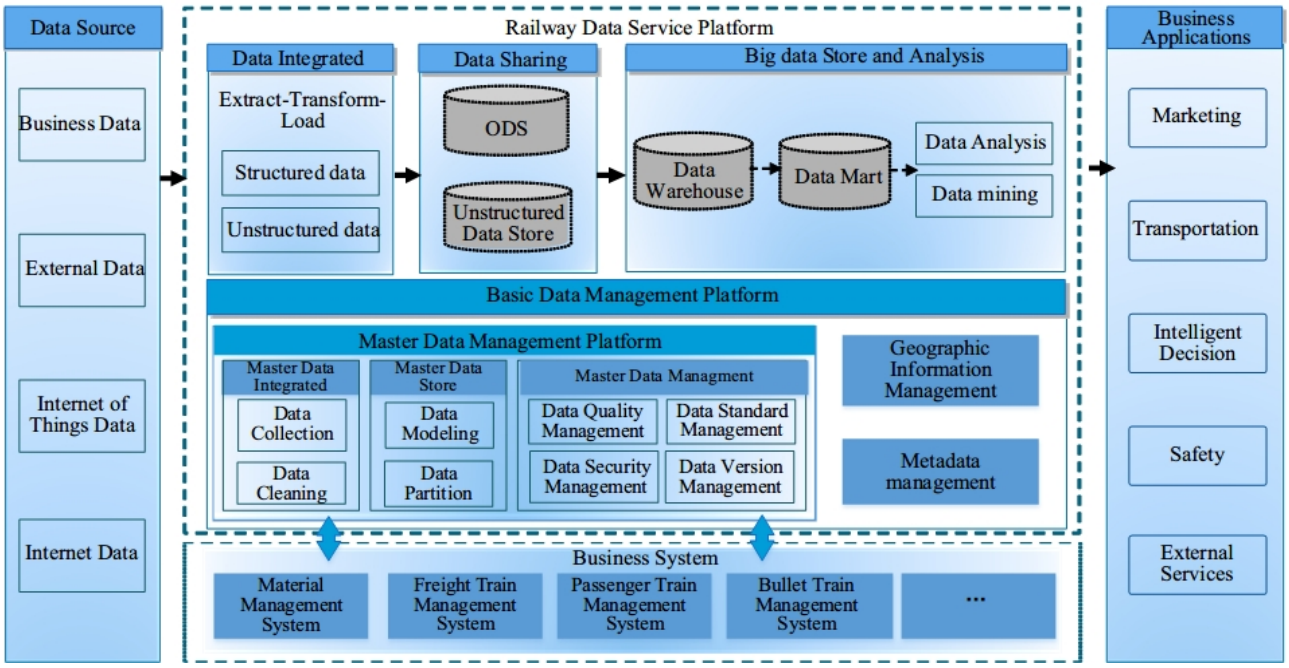

Fig.1 General Framework for Railway Master Data Management of Big Data Applications

Master Data Integration. As the railway master data is dispersed within each business system, the first step of master data management is to integrate the railway master data, which mainly includes the master data collection and the master data cleaning. Master data collection refers to the collection of master data scattered in various business systems to the unified railway master data management platform through the ways of the WebService interface, message queue, or importing Excel files. The data collection mainly includes two ways. One is entirely based on the national standard in which the content of master data is totally consistent with that of the standard. The other is the other is derived from some industrial standards and standard technical documents that may not be updated in time. In the practical application of the second way, the content may not be completely consistent with that of the standard. The platform adopts the collection scheme which is based on the content of standard and integrates with the business department to ensure the accuracy and authority of the data.

Railway master data cleaning refers to the identification and correction of dirty data in railway master data to improve the quality of master data. The problems about data quality are generally divided into three categories: null values, error data, and duplicate data. Null-value data and format errors can be filtered by regular expressions and fed back to the business department for modification. Duplicate data can be sorted and filtered by algorithms. , For the data with too many similarities, they should be judged manually.

Master Data Storage. The storage of railway master data adopts a relational database for data persistence. The integrated railway master data is uniformly modeled and then stored in the database according to the classification of the master data. In order to achieve a flexible management of metadata, the strategy of a master table plus a mapping table is adopted. By maintaining the metadata information of the master data separately in the mapping table and establishing a mapping relationship with the master table extended field of the actually stored data, the user can add, modify, or delete the attributes of the master data conveniently. In addition, with regard to the problem of excessive storage of single tables, a partitioned storage scheme is adopted for the master table which is performed according to the classification of the master data. Therefore the query of master data is effectively improved.

Master Data Management. Master data management is used to guarantee the instantaneity and high availability of enterprise-wide master data based on a set of constraints and methods, so that the core information of the enterprise can be reused and the core data of each application can be consistent [6]. Railway master data management mainly includes four parts: the management of data quality, data security, data standard, and data version.

(1) Quality Management of the Railway Master Data

Railway business systems have accumulated a large amount of business data, but due to the existence of data missing, data inconsistence, data duplication and other quality problems, the data cannot be effectively analyzed and utilized. Therefore, quality management of the data is a significant part of railway master data management. In the railway master data management platform, the quality of master data is controlled through three steps: the inspection, assessment and optimization of the data quality.

(2) Security Management of the Railway Master Data

Railway master data is the core basic data required by the normal operation of railway business information systems. During the use and maintenance of master data, the security requirements are extremely high. Any security problems caused by illegal alteration or leakage of master data will cause serious consequences on the transport and production of the railway [7]. Therefore, the platform adopts a variety of measures to ensure data security, which mainly includes three parts. First, the use of encrypted storage and regular backup mechanisms for master data content security; Second, the formulation of the authority control policies to refine the granularity of railway data fields and to ensure the security and maintenance of the access to the railway data; Third, the supply of a comprehensive function of log management, and automatic records of all the user's operations on the system, so the traceability of the platform can be ensured.

(3) Standard Management of the Railway Master Data

The railway master data standard is the precondition for realizing the integration of railway data. Only a master data standard is formed, can the quality of railway data be effectively improved. The railway master data management platform provides a coding standard module for the management of data standard documents. The module contains two parts: normal sort and normal data. The normal sort module mainly focuses on the management of standard classification system which makes organization and management of various standards more scientific and effective. The standards that platforms already have had can be divided into three categories: national standards, industrial standards, and standard technical documents. In the standard data module, administrators can add current standard documents to each standard category according to the classification, which will facilitate users' access and download and provide several basic functions such as standard addition, modification, deletion, and inquiry.

(4) Version Management of the Railway Master Data

The railway master data management platform provides the version management of the railway master data. After the master data is changed, the maintenance department of the platform can replace the original data with new data by adding the master data version and publish the latest railway master data standards to each service information system. At the same time, the data version management and traceability based on time stamp can retain all historical records of data on the basis of saving storage space so that the data's change track can be reviewed.

Railway Master Data Model Based on Knowledge Graph

The concept of the knowledge graph was originally developed by Google Inc. in 2012 to optimize the quality and efficiency of search engines returning query results. Knowledge graph refers to a network structure diagram formed by visual technology and interconnected by various points. It is mainly composed of node and relationship of nodes which is used to describe the relationship between resource entities. The nodes can be people, places, data, etc. which are also called entities. The entities can also include multiple attributes, which we call node relations. The knowledge graph can present structured topic information to provide the most comprehensive summary so that users can find the information they want.

The value that each isolated data provides is very limited. However, there is a natural connection between them. It can make better use of the data value to form a data asset graph by using knowledge graphs. The railway master data entities is modeled by the knowledge graph, and the railway master data model based on the knowledge map is shown in Figure 2.

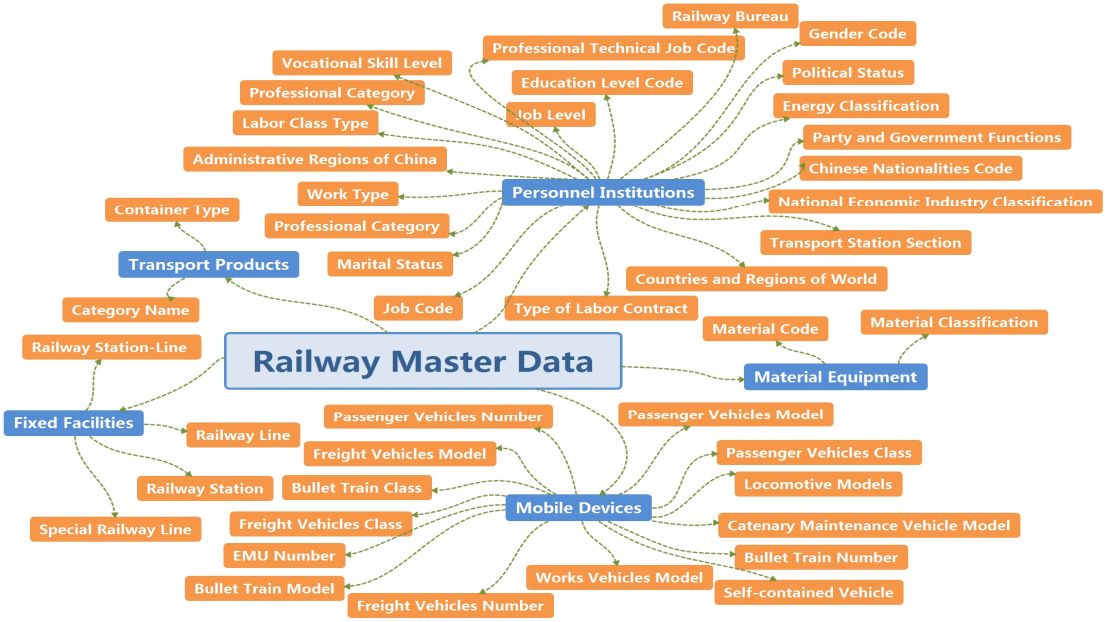

Fig.2 Railway Master Data Model Based on Knowledge Graph

By sorting out the entity and entity relationships contained in the railway master data, the railway master data is divided into five major categories: fixed facilities, mobile devices, material equipment, transportation products, and personnel institutions [8]. There are a total of 42 master data, and 192 fields included in the scope of railway master data management, which will continue to expand in the future. The category of fixed facilities includes station master data and line master data. The category of mobile devices includes master data of the vehicles such as passenger vehicles, freight vehicles and bullet trains, master data of the vehicle types and the vehicle numbers. The category of material equipment includes master data of material classification and material codes. The category of transportation product includes master data of names of products and types of the container box. The category of personnel agencies includes master data of railway industry such as master data of railway bureau and transport station section, and national standard master data such as codes of professional technical posts, energy classification, administrative division, and economic industrial classification.

In the entire process of master data modeling, there are mainly four steps. The first is knowledge acquisition, extracting the master data types and fields from the business data, and solving the decomposition and sequencing at the level of business. The extracted knowledge is then expressed by using a triplet of entity-relationship-entity. Afterwards, the traditional RDF (Resource Description Framework) storage or graph database is used to store the identified knowledge, and the visualization tool is used to display the knowledge graph finally.

Key Technologies for Railway Master Data Management

Rule-Based Data Loading and Cleaning Technology. The platform innovatively uses rule-based data loading and cleaning technology to solve the problems of Initialize data import of master data. The technology realizes the modeling and initialization of data by pre-specifying the rules of the encoded data, and formulating the cleaning scheme to clean the data imported into the platform. So that the imported data can meet the requirements of the platform data management. The dirty data among the raw data can also be found in time and further feedback and modification will be conducted.

The data import process of the platform is based on the idea of data modeling of master data. Data rules are constructed by using attribute definitions and encoding rules of the master data. Data types and encoding rules of corresponding master data attributes are predefined in the rules, and different rules are built for different master data to meet the individual needs. Rule-based data loading and cleaning process is shown in Figure 3.

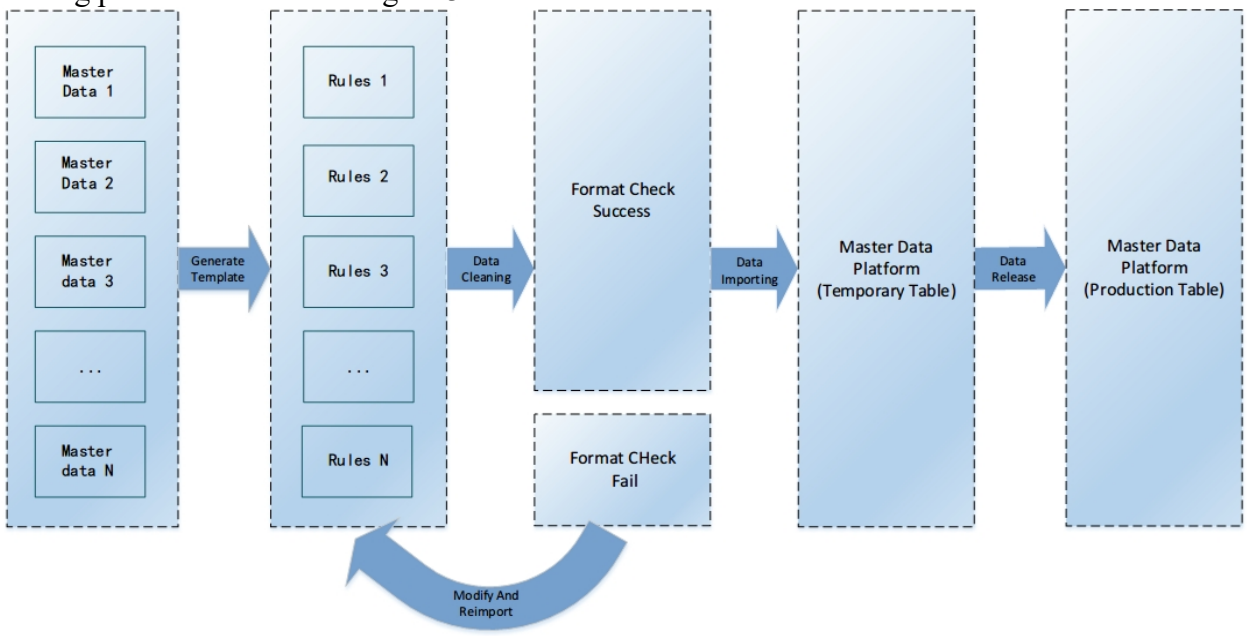

Fig.3 Rule-Based Data Loading and Cleaning Technology

First, model the master data and generate a template. Before importing data, the user needs to model the master data on the platform, including creating master data, adding fields of master data, establishing rules for the contents of the fields and configuring other attributes. The user then needs to download the data rules that have been modeled and fill the data in the field format required by the data rule. After filling, the user needs to upload the data file, fill in the detailed information about the master data, and clean the data to be imported. Data cleaning verifies all the records in the data file according to the requirements of the rules. If the data meet the requirements, they will be cleaned. While if the data type or encoding rules are not satisfied, the records are marked as dirty data and returned to the user. After the modification, the user can modify the data into the correct format and clean the data again. Finally, the data passed through the check will be loaded into the temporary database, but the data has not yet been officially released to the public. After the data administrator reconfirms the data and releases it, the data can officially enter the production database and be used by other business systems. So far, the data loading and cleaning process comes to the end.

Timestamp Based Data Version Management Technology. The technology of data version management based on the timestamps is to record the moment that the data changes to identify the changing data, then complete records of the entire process of data field changes. The efficient storage and query can be achieved finally. Railway data has its specificity compared with other industry data, such as the name change of a station, the alteration of status and so on. During the process of the statistical analysis and calculation of the big data, it is required to record the history of data changes to obtain a complete evolution of the data. If all versions of data are stored in full, especially for master data with a large amount, the pressure of database will increase exponentially as the amount of data increases. Therefore, in the design and construction of the platform, data version management and traceability techniques based on timestamps are used to achieve both the full version retention of data and the reduction of database pressure.

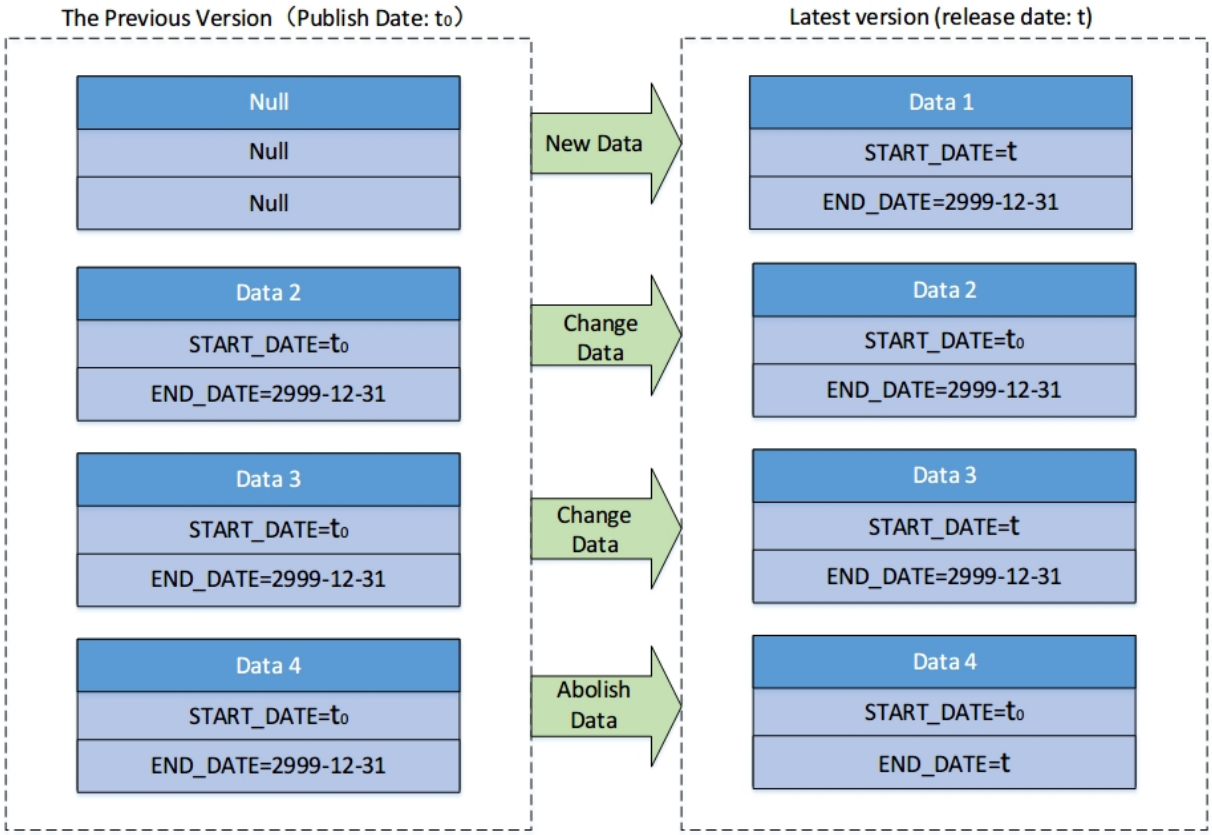

Fig.4 Timestamp Based Data Version Management Technology

The update of the master data generally includes three types: addition, modification, and revocation. When designing the structure of the database table, a main table and a historical attributes change table are established for the storage of the master data, and both the effective date and the

expiration date are set to record changes in the data between different versions so that the full data and incremental data of any data version can be extracted. As shown in Figure 4, when data is newly added, the effective date of the newly added data, that is, START_DATE is set as the release date of the current data version, and the expiration date, that is, END_DATE is set as the maximum value of the date (2999-12-31). When the data changes, the effective date of the changed data is set as the release date of the current data version, while the abolition date is not changed and the data attribute before the change is stored in the historical attribute change table. The effective date is set as the date before the data change while the date of expiration is set as the release date of the current version of the data. For the retired data, its effective date does not change, and the expiration date is set as the release date of the current version. When any version of the data is queried, the full data of any version can be found by filtering the condition "START_DATE<= release date of this version <END_DATE" and combining the data records in the main table and the historical attributes' change table.

The Application of Railway Master Data Standard to the Big Data Integration

The railway data service platform adopts the StreamSets to import business data, and can collect and process structured data and unstructured data. The entire process from data source to platform storage can be realized through configuration, development and process monitoring. StreamSets is a lightweight and powerful engine that simplifies the process of data collection to the maximum extent. It can extract, convert, load, and complete data collection in the simplest way. StreamSets includes support for hundreds of different components. The data extraction function supports bulk import and real-time synchronization. The bulk import can be collected from databases such as Oracle, MySQL, and MongoDB through the JDBC component. The real-time synchronization function supports models such as Kafka and Flume. And Oracle CDC and other modes; data conversion function can be a specific field of data type conversion, removal, merging, segmentation, replacement and sorting and other operations; data loading supports full loading and incremental loading, which is convenient for users to select the most appropriate loading method and can satisfy the requirements of various business situations to the greatest extent. StreamSets can also monitor and manage the process of data collection in real time, automatically detect the data that do not meet the configuration rules, and output the wrong information in real time.

At present, the data service platform has collected data of 16 systems and 8 departments including passenger tickets, bullet trains, traffic safety and infrastructure inspections, more than 160 data tables and 3,200 fields. However, due to the large number of master data fields included in the data of each business system, for example, the station field in ticket stub data and the station field in the EMU system may be inconsistent. As a result, the associated analysis of the subsequent data cannot be performed through master data. Therefore, after the data is collected from the business system to the source area through the StreamSets, a mapping relationship is established between the field names of the service data and the master data fields, and then the corresponding master data rule is loaded to clean and standardize the business data. The standardized data are provided for the analysis of subsequent big data. The application process of the master data standard in big data integration is shown in Figure 5.

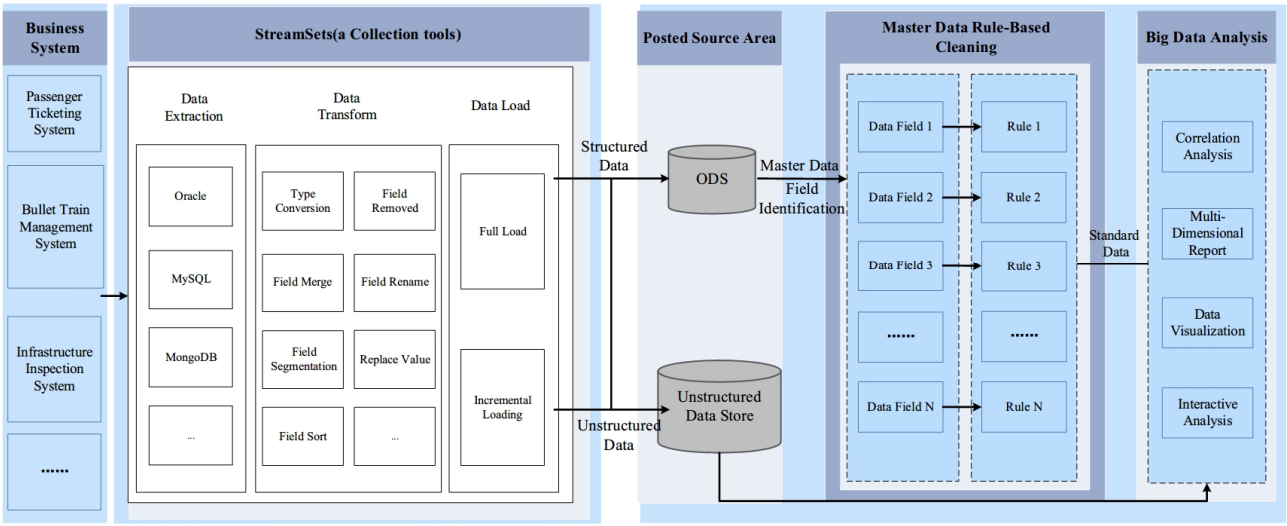

Fig.5 the Application of Railway Master Data Standard to the Big Data Integration

Conclusions

In order to meet the needs of analysis and application on railway big data, this paper proposes a framework of railway master data management in terms of big data applications, covering master data integration, storage and management, and establishing a master data model based on knowledge graphs. With the application of some key technologies such as data cleaning and data version management of railway master data, the effective management of railway master data can be achieved. Finally, taking the application of the railway master data standard in the railway big data integration as an example, using the railway mater data standard to clean the service data of the railway data service platform will greatly improve the data's quality of the railway data service platform and lay a solid foundation for the data analysis and application.

References

[1] WANG Tong-jun. On Top-Level Design for China Railway's Big Data Application & Case Study [J].China Railway, 2017(01):8-16.

[2] DAI Mingrui, ZHU Kefei, ZHENG Pingbiao. Thoughts on Applying Big Data Technology to China Railways [J]. Railway Transport and Economy, 2014, 36(03):23-26.

[3] China Academy of Railway Sciences. General Plan for Railway Data Service Platform [R]. Beijing : China Academy of Railway Sciences,2017.

[4] China Academy of Railway Sciences. General plan for railway public infrastructure coding and master data management platform [R]. Beijing : China Academy of Railway Sciences,2017.

[5] MA Xiaoning, ZOU Dan, WU Yanhua. Solution and Application of Railway Master Data Management Platform [J].China Railway,2017(01):17-23.

[6] ZHAO Fei. Master Data Management Based on Whole Lifecycle: Detailed Explanation and Practice of MDM [M]. Beijing: Tsinghua University Press,2016.

[7] LI Ping, ZHAO Bing, LIU Yi-fei. Study on Railway Big Data Security Technology System for Full Life Cycle [J].China Railway, 2018(02):32-36.

[8] ZOU Dan. Key Technologies of Railway Master Data Management Platform[J]. Railway Computer Application,2017,26(01):31-35.