ATLANTIS PRESS

Advances in Engineering Research, volume 150

4th International Conference on Machinery, Materials and Computer (MACMC 2017)

Study on the Data Acquisition and Cleaning Technology in Mixed Heterogeneous System

Li Wencui[1,a], Dong Gang-song[1],Li Xiong[1],Tang Weixia[2], Zhang Yong[1], Yang Yi[1]

[1]Information&Telecommunication Co. of State Grid Henan Electric Power Company, Zhengzhou, 450052, China

[2]China Information Technology Designing Consulting Institute Co. Ltd., Zhengzhou, 450000, China

[a]email:elf8650@163.com

Keyword: big data, data acquisition, data cleaning

Abstract. Big data, as the frontier technology of data analysis, can quickly obtain valuable information from various types of data. This article is based on unified, standard data collection platform. It collects multi-word configuration, alarm and performance data. Combined with a dumb associated resource data, and research the data acquisition clean governance. Thus, the quality of data sets can be improved to meet the demand of data analysis.

Introduction

With the advent of the cloud era, Big data has attracted more and more attention. Compared with the traditional data warehouse application, it has the characteristics of large data volume and complex query analysis.

Big data, as the frontier technology of data analysis, can quickly obtain valuable information from various types of data. This article is based on unified, standard data collection platform. It collects multi-vendor configuration, alarm and performance data. Combined with a dumb associated resource data, and research the data acquisition clean governance.

The data and correlation data of terminal communication access network are evaluated in big data to achieve the accuracy, reliability and completeness of data. It can convert "incomplete data" into data that meets data quality requirements or application requirements. Thus, the quality of data sets can be improved to meet the demand of data analysis [1-2].

Data Acquisition and Governance

The Intelligent Acquisition of Dumb Resource. The intelligent acquisition of dumb resource can also be called as unstructured data intelligent access, which can generate standard data that is general, clear and correct by standardized form technology and intelligent data processing, and then is imported by the methods (as shown in Tab.1).

The Cross-Professional Data Acquisition. The cross-professional data acquisition is to obtain data information from other network management systems (such as TMS state grid communication management system, power distribution automation open-3200 system, production management PMIS system) by standard Web-Service interface system integration with a third party based on XML with the third party system, which transmit date by way of Web-Service [3-4].

The cross-professional data acquisition can query the medium and small capacity data through a direct call other professional system of Web service interface. The other professional system can query the data at the corresponding level access network management system by calling the corresponding access network management system of the Web service interface.

By adopting the Web and TCP method, the system meets the requirements of real-time, transmission efficiency, openness and low maintenance difficulty (as sown in Fig.1 and Fig.2).

Tab.1 Unstructured data interface specification

Num	The name of the interface	The interface way	Interface classification
1	Add the document	WebService/HTTP	Document management
2	Add documents (including format files)	WebService/HTTP	Document management
3	Add formatting files	WebService/HTTP	Document management
4	Batch add documents	WebService	Document management
5	Modify the document	WebService/HTTP	Document management
6	Download the document	WebService/HTTP	Document management
7	Download the format file	WebService/HTTP	Document management
8	Delete the document	WebService/HTTP	Document management
9	Get document properties	WebService/HTTP	Document management
10	Document links to the new directory	WebService	Document management
11	Mobile document	WebService	Document management
12	Query document list	WebService/HTTP	Document management
13	Add documentation vertically	WebService/HTTP	Longitudinal exchange
14	Download the document vertically	WebService/HTTP	Longitudinal exchange
15	Document properties are obtained vertically	WebService/HTTP	Longitudinal exchange
16	Vertical sync document	WebService/HTTP	Longitudinal exchange
17	Add distribution documents	WebService/HTTP	Document distribution
18	The distribution document is coded by policy	WebService/HTTP	Document distribution
19	Distribution of the document	WebService/HTTP	Document distribution
20	Query the distribution document list	WebService/HTTP	Document distribution
21	Determine the folder presence	WebService	Folder management
22	Gets the subfolders under the folder	WebService	Folder management
23	Gets the subfolders under the folder	WebService	Folder management
24	Online browsing	JS package	Online browsing

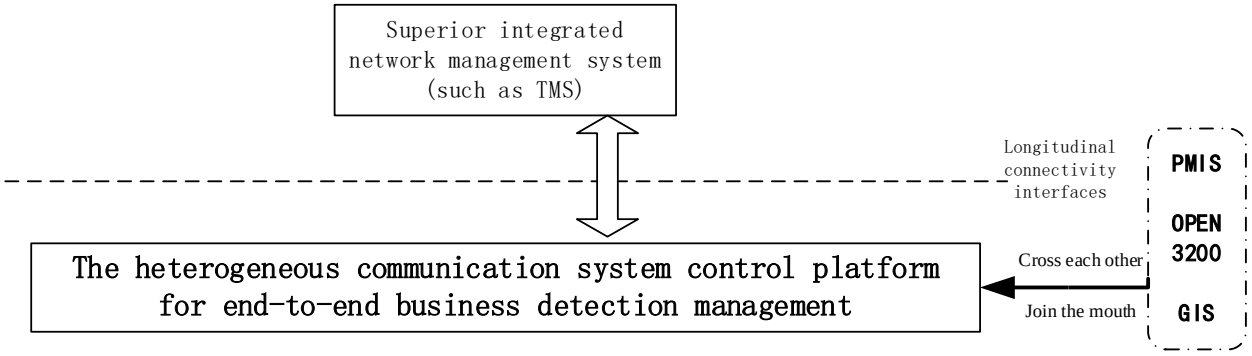

Fig.1 Deployment architecture

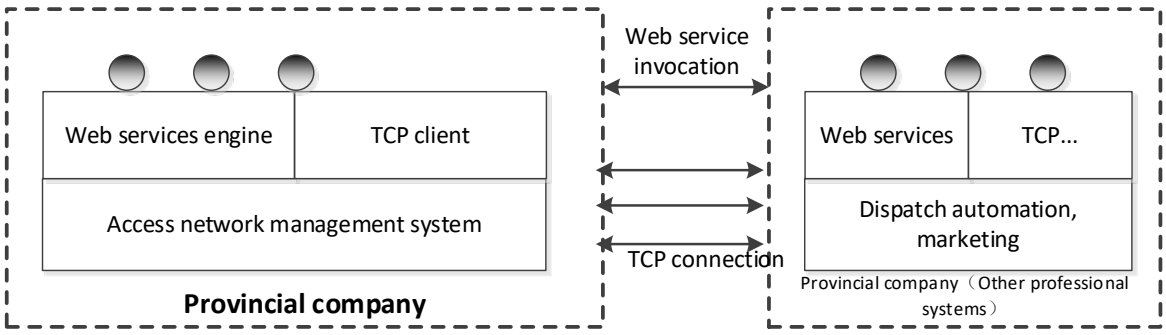

Fig.2 Logical architecture

Large Data Acquisition and Cleaning Evaluation

Dataeye Data Cleaning Process. Data cleaning process is shown in figure 3: such as filling missing data, eliminate noise data, and so on, mainly through analysis of the causes of "dirty data" and the existing form, using existing data mining means and methods to clean "dirty data", "dirty data" can be converted to meet the requirements of data quality requirements or application data, so as to improve the quality of the data set, meet the needs of present data analysis.

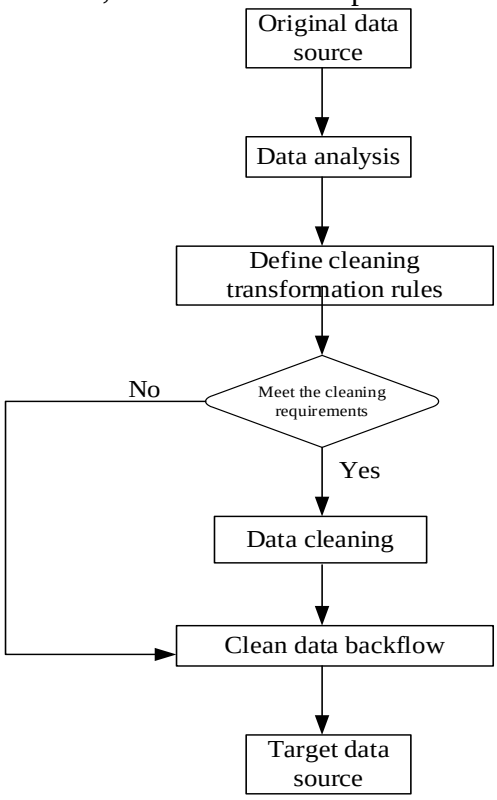

Fig.3 DataEye data cleaning process

Dataeye Data Cleaning Solutions and Practices. The data cleaning scheme is shown in figure 4: Using techniques such as mathematical statistics, data mining or pre-defined cleansing rules to translate dirty data into data that meets data quality requirements. Data cleaning can be used to deal with data loss, transboundary value, inconsistency code, and duplicate data in terms of accuracy, completeness, consistency, uniqueness, timing and effectiveness of data.

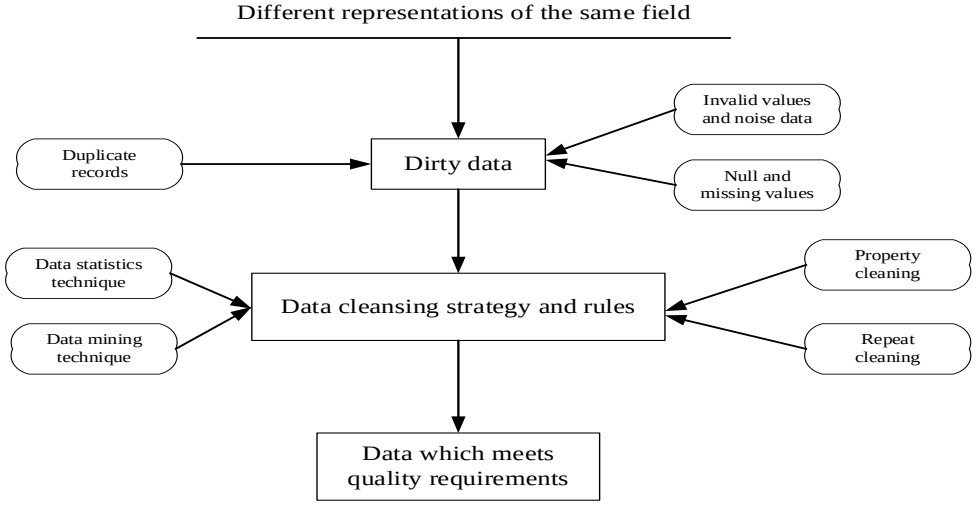

Fig.4 DataEye data cleaning solution

Effective Cleaning Method. The main task of data cleaning is to detect and remove/correct the dirty data that will be loaded into the data warehouse. Multiple heterogeneous data sources and

massive data cleansing need to be integrated with data extraction and data conversion. It is used in uniform with data and needs to be recycled. If the data source is a smart network pipe collection data (see data source 1 and data source 2 in figure 5), it can use SQL to do part of the data cleaning in the data extraction process. But there are some unstructured data sources (such as data source 3 unstructured data) that can be extracted directly from the data source. Then clean it when the data is converted[5-6]. Data cleaning in data warehouse is mainly carried out during data conversion.

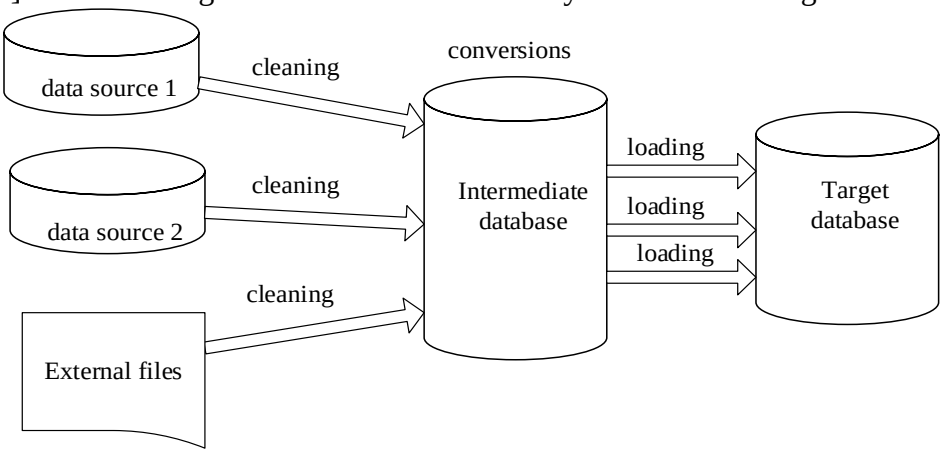

Fig.5 Data cleaning model

Conclusion

Big data, as the frontier technology of data analysis, can quickly obtain valuable information from various types of data. This article is based on unified, standard data collection platform. It collects multi-word configuration, alarm and performance data. Combined with a dumb associated resource data, and research the data acquisition clean governance. Thus, the quality of data sets can be improved to meet the demand of data analysis.

References

[1] Xue Kui, He Shi-lin. The Research of Electric Information and Communication Risk Management [J]. Communication Technology, 2014,8.

[2] Li Wencui, Li Xiong, Zhang Chi, Wu Lijie, Shu Xinjian, Tang Weixia. Research on the Figure Module Integration Technology in Communication Monitoring. ICEEECS, 2016(50), 686-689.

[3] Li Meng-xing, Chi Cheng-zhe, Wang Hai-yan. The Trend and Preventive Measures of Electric Information Safety Management [J], Power Information, 2010(8),12.

[4] Li Wencui, Li Xiong, Yang Ying, Ding Ying, Shu Xinjian, Zhang Yong. A Brief Talk on Information and Communication Safety Management of Electric Power Enterprise. ICMMBE, 2016(83), 217-220.

[5] Wang Ji-ye, Guo Jing-hong, Cao Jun-wei, Gao Ling-chao, Hu Zi-wei. Review on Information and Communication Key Technologies of Energy Internet [J]. Smart Grid, 2015.

[6] Li Wencui, Shu xinjian, Li Xiong, Gao Hui, Liu Bo, Wang Chunying, Yang Ying. An Information Security Management System Based on “Five-in-one”. ICEEECS, 2016(50), 743-745.