Research on Acquisition of Clean Governance Evaluation Techniques for Big Data

Wu Li-jie, LI Wen-cui, WANG Chun-ying, Yang Yi, An Zhi-yuan, Zhang Ning-ning

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

374663965@qq.com

Keywords: Large Data; Clean; Governance; Incomplete data; Quality

Abstract. In this paper, based on large data for clean governance evaluation technology, the terminal communication access their own data and the correlation data for big data intelligent access to cleaning, warehousing management assessment, implementation based on RCM/RBM and LCC communication equipment state model, state evaluation and auxiliary decision-making, “missing or incomplete data” can be converted to meet the quality requirements or the requirements of the application data, data communication terminal access their own data and associated data cleaning governance evaluation research, communication large data integration management model is put forward, build communication terminal access network data platform, so as to improve the quality of the data set, meet the needs of data analysis.

Introduction

After recent years of electric distribution network informatization, accumulated the massive amounts of raw data, including alarm equipment, performance and configuration data and operation maintenance data and professional management, etc. The intelligent device resources can obtain relevant data through the factory's network pipe interface, and ensure the accuracy, reliability and completeness of the data of intelligent equipment resources [1]; but Dumb resources(Dumb resource devices and Dumb resource connections) as an integral part of big data analysis, "Dumbresources" don’t open his mouth, big data analysis difficult, so dumb data through large data intelligent access to resources, cleaning and management technology, can better for advanced application provides important basis for later data analysis.

Domestic and Foreign Research Level

Domestic and Foreign Research Status. With the rapid development of information technology, every field in every moment at an alarming rate to produce all kinds of huge data information, humans also in all aspects of the work life come into contact with more and more data information. Humans, however, the lack of understanding of data information and the trend of the data explosion is not symmetrical, human is trying to convert data into favorable information knowledge at the same time, also faces big data with "dirty data" challenge, in the original data source management, cleaning is transformed into can be utilized to understand the target data source, become the human understanding is important step in the process of data.

With the advent of the era of big data, the consciousness of "data driven operations" in the game industry gradually popularization, "data driven by refinement operation" has become inevitable trend in the process of game operations, but also in facing the challenge of "dirty data". DataEye, as a professional third-party game data analysis service provider, will combine the experience of actual work to build the DataEye data cleaning and management system.

Research Status of Power Industry. At the end of 2011, the communication management system carried out the data of the dumb resource data of power communication network. After the communication management system deployment, resource data entry into the system database, the traditional way of data entry, using the data of static form, graphics by professional personnel, according to their own grasp of the communication network and the cognitive system data entry, manual dumb resource data communication management system inventory check entry work, and then by site personnel in accordance with the system in the implementation of input data to complete the connection of the network resources or configuration, etc.

In 2012, the information communication branch of Jiangsu electric power company carried out the research and development of power communication network channel serial algorithm, researchers at the beginning of the work [2], in view of the communication network of the most important business necessary parameters were analyzed, and clear through the north to the parameters of the interface is available for business, at the same time, the network structure of the network and has the business analysis, determine the crossover, topological connection, the basic information, such as optical fiber connection is subject to the above information for the channel path concatenated original algorithm, and applies the algorithm to the communication management system of intelligent design function and the function of resource scheduling in the channel. The application of intelligent channel design function and resource scheduling function lays the foundation for future research work.

Big Data Base Theory and Technical Route

Big data technology refers to the rapid acquisition of valuable information from a variety of large amounts of data. Big data technology is central to solving big data problems. It can be divided into eight categories: data acquisition, data access, infrastructure, data processing, statistical analysis, data mining, model prediction, and result presentation. Big data technology mainly forms the three calculation modes: batch processing, flow processing and interaction analysis.

(1)Batch Processing technology is represented by MapReduce and Hadoop system.

(2)Stream Processing technology is represented by Yahoo's S4 system and Twitter's Storm system.

(3)Interactive Analysis technology is represented by the Dremel system of Google.

Big data is not only a huge amount of information, but also the human's screening and processing of information. Big data processing methods have many, generally applicable big data processing flow, can be summarized as four steps, are collection, import and preprocessing, statistics and analysis, and finally data mining.

Acquisition. Big data acquisition refers to the use of multiple databases to receive from the client (Web, App or sensor forms, etc.), and the user can through these database for simple queries and processing work. For example, the e-chamber uses traditional relational databases MySQL and Oracle to store each transaction data, and in addition, NoSQL databases such as Redis and MongoDB are also used for data collection.

In large data collection process, its main characteristic and challenge is high concurrency, because at the same time may have tens of thousands of users to access and manipulate, such as train ticketing website and Taobao, they reached millions of concurrent traffic during peak, so need to deploy a large number of database on the acquisition end to support. And how to load balancing and sharding between these databases requires deep thinking and design.

Import/Pretreatment. Although there will be a lot of database on the collection side, but to the analysis of these huge amounts of data efficiently, or should the data import from front end to a centralized large distributed database, or distributed storage cluster [3], and can be based on the import do some simple cleaning and pretreatment. There are also some users who use the Storm from Twitter to do a flow calculation on the data to meet the real-time computing requirements of some businesses.

The characteristics and challenges of importing and preprocessing are mainly imported data volumes, and the number of imports per second can often reach hundreds of megabytes or even gigabit levels.

Statistical Analysis. Statistics and analysis the main use of distributed database, or distributed computing cluster to store huge amounts of data in its ordinary analysis and classification summary, etc., in order to satisfy the demands of most common analysis, in this regard, some real-time demand would use the EMC GreenPlum, Oracle Exadata does, as well as the column type based on MySQL storage Infobright, and some of the batch, or demand can use Hadoop based on semi-structured data.

The main characteristics and challenges of statistics and analysis are the large amount of data that is involved in the analysis, which can be of great use to system resources, especially I/O.

Mining. As the previous statistics and analysis, data mining, and laborers typically have little predefined theme, mainly on the existing data calculation, based on all kinds of algorithm to Predict effect, so as to realize some high level data analysis needs. The typical algorithms have Kmeans for clustering, SVM for statistical learning, and NaiveBayes for classification [4]. The main tools used are Mahout of Hadoop.

The characteristics and challenges of this process are mainly used to excavate the algorithm is very complex, and the calculation involved data volume and calculation amount are very big, the common data mining algorithm is mainly single-threaded.

A widely accepted processing model is the multi-processing phase model designed by Fayyad et al.

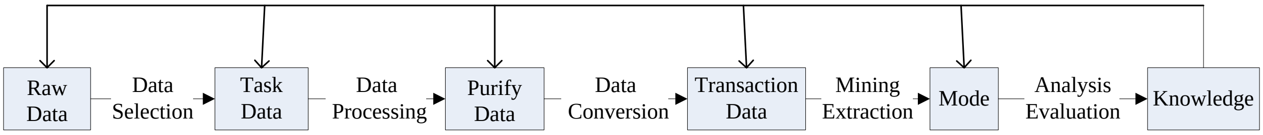

Fig.1 the multi-processing phase model

At present, big data research mainly as a research method or a tool of discovering new knowledge, rather than the data itself as the research target, it is closely related with the traditional data mining methods have radically different.

Large Data Acquisition And Cleaning Evaluation

DataEye Data Cleaning Process. Data cleaning: 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.

Repeat Value Cleaning Governance. In a complex work environment, the occurrence of repeated values of data is common due to multiple reports of data or other human factors, which mainly uses the field similarity to identify the repeated values.

Field similarity definition: the similarity between fields is calculated according to the contents of two fields, and the value of the similarity degree of two fields is calculated, O<S< 1.The smaller the S, the higher the similarity of the two fields; If S=0, the two fields are the full repeating fields [5].Depending on the type of field, the calculation method is different.

Boolean field similarity calculation method: for Boolean field, if two fields are equal, then the similarity is equal to 0. If different, the similarity is equal to 1.

Research on the Method of Accurate Estimation of Big Data

Data evaluation is short for "data quality assessment", from the Angle of the integrated application of data, information and data collection, storage, and output to conduct a comprehensive inspection and evaluation, thus, improve the credibility and effectiveness of the information and data, providing a better basis for decision making. It is different from the common sense of the quality evaluation, but on the application of the data from the enterprise point of view, from the perspective of enterprise management requirements, data to carry on the deep analysis of the enterprise, again to the flow of information to make the necessary adjustment, to adapt to the actual needs of enterprise management, and not just as easy to ensure accurate data.

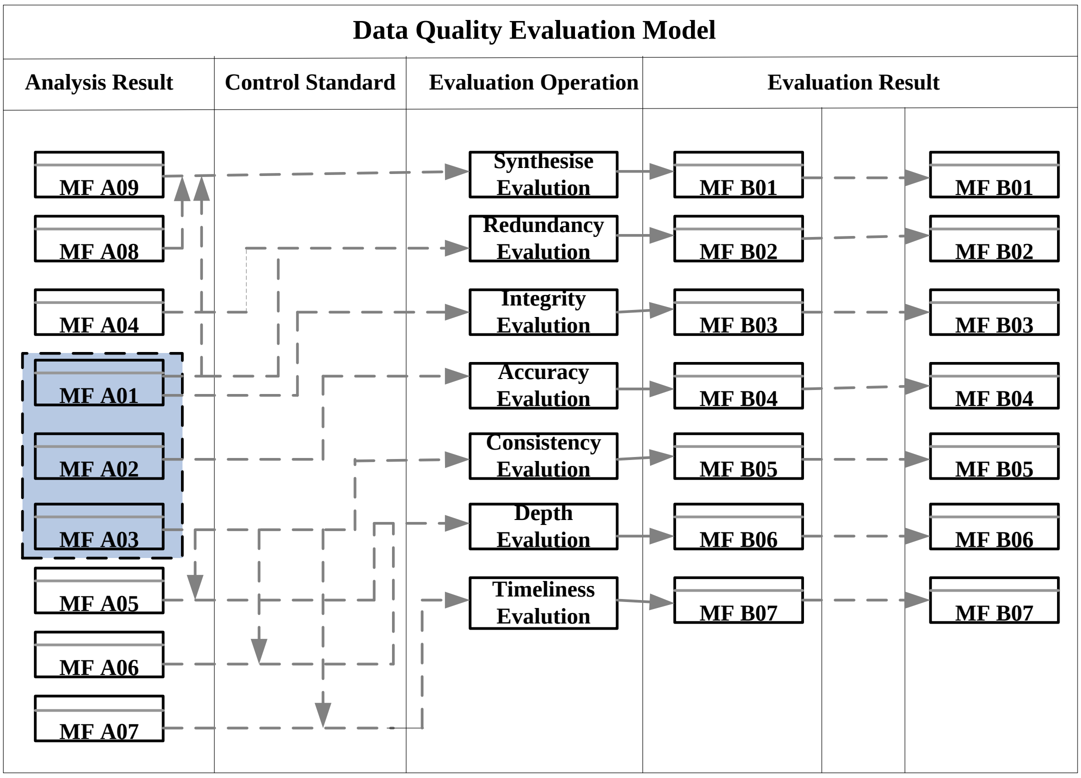

Fig.2 the data quality assessment model

Conclusion

Terminal communication access network big data platform is to save the company management system of city function extended data acquisition platform, should be in accordance with the requirements of the relevant standards, has been general, clear [6], correct format added province communication management system for city data acquisition platform model structure, complete the communication terminal access to the underlying data of intelligent acquisition, effective cleaning, strict management and accurate evaluation.

References

[1] WEI Yu-fan. Design and Realization of Equipment Maintenance Information Management and Analysis System[J].Computer and Information Technology, 2009,17(4):62-66

[2] Li Wen-cui. An Information Security Management System Based on “Five-in-one”. ICEEECS, 2016(50),743-745.

[3] YU Da-wei, ZHANG Mao-qing, LI Qiang, SUN Chang-ling. Design of Remote Monitoring System Based on Global System for Mobile Communication Network [J]. Electro Technics Electric, 2009(11): 25-28

[4] Li Wen-cui. Research on the Figure Module Integration Technology in Communication Monitoring. ICEEECS, 2016(50), 686-689.

[5] HE Li, WEI Xue-wen, HAN Huan-ju. Security of Business Information & Network Resources in the Third Generation Mobile Communication System [J].China New Telecommunications, 2013(14): 1-4

[6] Li Wen-cui, Li Xiong. A Brief Talk on Information and Communication Safety Management of Electric Power Enterprise. ICMMBE, 2016(83),217-220.