Data Quality Assessment for On-line Monitoring and Measuring System of Power Quality Based on Big Data and Data Provenance Theory

Tian Hongxun, Wang Honggang
State Grid Corporation of China
No. 86, West Chang'an Avenue Xi A
Beijing, China

Zhou Kun
State Grid Anhui Electric Power Co. Ltd
No. 9 Mount Huangshan Road
Hefei, China

Shi Mingtai, Li Haosong, Xu Zhongping, Kang Taifeng, Li Jin, Cai Yaqi
Beijing State Grid Xintong Accenture Information Technology Co. Ltd
No. 29, North Xinhua Street
Beijing, China
e-mail: yayakero@126.com

Abstract—**Currently, on-line monitoring and measuring system of power quality has accumulated a huge amount of data. In the age of big data, those data integrated from various systems will face big data application problems. This paper proposes a data quality assessment system method for on-line monitoring and measuring system of power quality based on big data and data provenance to assess integrity, redundancy, accuracy, timeliness, intelligence and consistency of data set and single data. Specific assessment rule which conforms to the situation of on-line monitoring and measuring system of power quality will be devised to found data quality problems. Thus it will provide strong data support for big data application of power quality.**

Keywords-power qualiy; data quality; big data; data provenance; data assessment

I. INTRODUCTION

Power quality data has the richest indicators and the finest granularity among various kinds of grid monitoring data. The System accumulated huge amount of history data during its long-time running. Meanwhile, the on-line monitoring and measuring system of power quality integrates with such as production management system and power utilization collection system, that will produce more multi-source and isomerous data. In the age of big data, there are three dimensions to big data known as volume, variety and velocity, therefore, high-quality data is the precondition for big data to play its role [1].

Data quality assessment based on big data has multiple dimensions, therefore many scholars consider that big data's data quality has high requirement for consistency, integrity and usability [2]-[3]. At present, data quality assessment about electrical power mostly identify evaluation object, choose evaluation indicator, devise rule, decide weight and then calculate a score to construct a data quality assessment model [4]-[5]. Some scholar assess data quality problems based on cloud model, abnormal value checkout method [6]. Or combine analytic hierarchy process and grey cluster method to comprehensively evaluate data quality of power quality [7]. Those methods don't relate to the grid's characteristic of cross-system, cross-business and trans-department closely.

Therefore, this paper will combine data quality assessment with real power business which has the huge amount of data, a wide variety of equipment, complex business system. Use data provenance theory to assess the dimension of redundancy, accuracy, consistency, Timeliness and intelligence of data quality. Design the checkout rule that fit to the specific business and study the source of data quality problems from every link of the system. That will normatively process the basic different kinds of data from different source, unified storage and integrate data.

II. ON-LINE MONITORING AND MEASURING SYSTEM OF POWER QUALITY

The power quality on-line monitoring and measuring system of State Grid Corporation of China is the largest data management platform of power quality at home and abroad at present. The aim of the system is to meet the demand of the business analysis and application for power quality indicators. These indicators include the grid frequency, grid voltage and reliability indicators. They refer to data integration of various application system include production management system, power supply voltage automatic collection system, dispatch technical support system and power utilization collection system. The functional diagram of on-line monitoring and measuring system of power quality is as Fig. 1. The data of those systems is pushed to data center, and then collected to the main power station, at last, it is uploaded to the headquarter. The data integration flow is as Fig. 2.

With the enlarge of the grid's scale and the increase of the data category, coupled with long accumulation of time, the data amount collected in the system grows linearly as well. That made the on-line power quality monitoring and

978-1-5386-4301-3/18/$31.00 ©2018 IEEE

measuring system requires higher data handling capacity, storage capacity and statistical analysis capability [9].

Power Quality Business is mainly divided into three business segment, those are voltage management, frequency management and reliability monitoring and measuring management. Every segment has its own indicators to monitor and manage power quality. Those indicators support seven class basic data include grid voltage, grid frequency, power supply voltage, power transmission and transformation reliability basic account, power transmission and transformation reliability operation event, power supply reliability basic account and power supply reliability operation event.

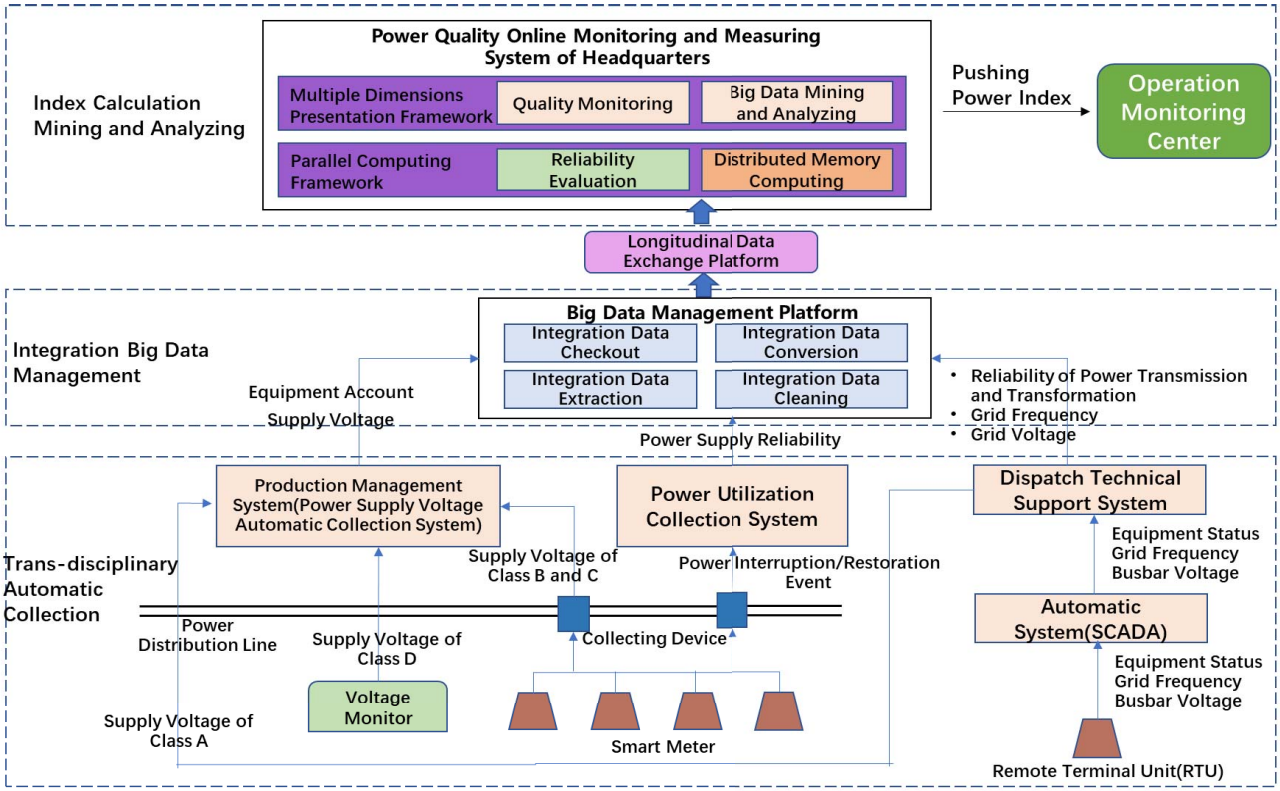

Figure 1. Functional diagram of on-line monitoring and measuring of power quality[8]

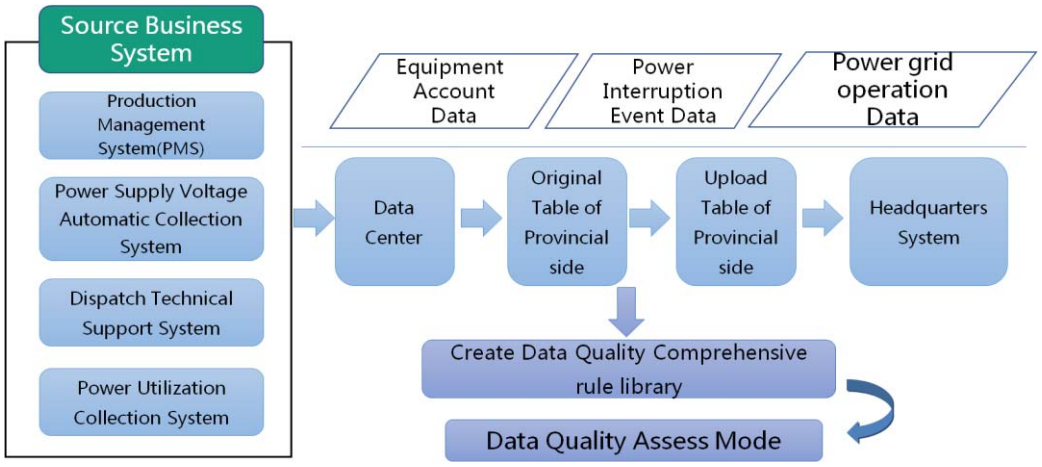

Figure 2. The data in on-line monitoring and measuring of power quality integration flow

III. Data Quality Assess System Based on Data Provenance

Data provenance designed to show how data transmitted from a data source to the data system through processing and has current data form. Data provenance is about to retrospect data source and gain evolutionary process of data by data provenance trace. This method can measure reliability and quality of data. Currently, there are two main methods of

data provenance trace, those are tagging method and reverse query method [10]-[11].

The data provenance the earliest is only for the database, data warehouse system, and later developed into various areas of the relatively high data authenticity requirements: such as biology, history, archaeology, astronomy, medicine and so on. With the rapid development of the Internet and the frequent network fraud, more and more people doubt the authenticity of the data, the data the authenticity of the increasingly high demand. Data traceability has become an effective way to research data and research data, set off a wave of traceability boom, therefore, data tracking computer gradually extended to all walks of life

Among them, tagging method is widely used. It traces the historical status of data through recording and processing relative information. In the on-line monitoring and measuring system of power quality we could also use the similar method which can consider data key field or business key field as tagging information of data to obtain the information of data provenance. Combing with big data theory and data quality assessment research results, data quality can be assessed in different dimensions like integrity, accuracy, consistency, uniqueness and so on. Design data quality assessment method for online monitoring and measuring system of power quality, tease the relationship between power quality data and construct data quality assess dimensions based on data provenance.

A. *Construct Data Quality Analytic Model*

This paper looks for data characteristics by tracing every link of data transmission and analyze these characteristics combining with tagging method. The analysis mainly concentrates on tans-disciplinary and multisystem integration of power quality data. Combining existing data-assess research results, we choose intelligence, redundancy, integrity, accuracy, consistency and timeliness to assess data quality of power quality. The data quality assessing rule model is as Fig. 3.

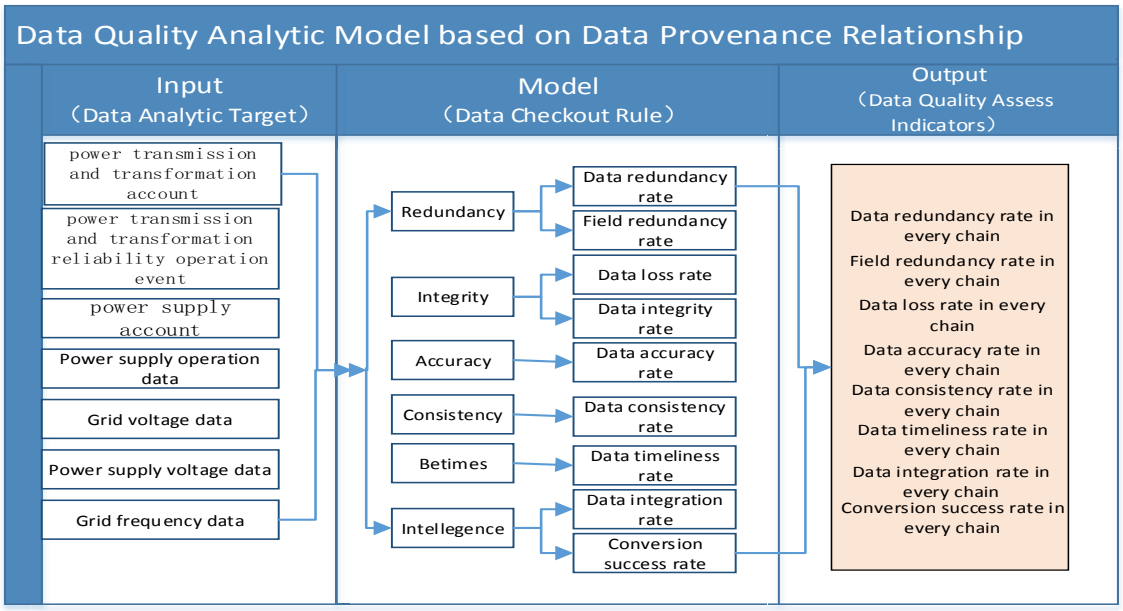

Figure 3. Data quality analytic model

B. *Design Data Quality Assesss Rule Set*

According to the six dimensions of redundancy, integrity, timeliness, consistency, intelligence and accuracy, we design corresponding evaluation rules that conform to actual assessment objects, form a set of data quality comprehensive check rules, and build data quality assessment model. Table I shows the rules for the calculation of specific indicators.

The check rule set divides the checksum into data sets and single data. Data set is the check of the total amount of data in the process of data transfer, and the data range is the verification of field change and logical relationship in the process of single data transfer. When doing the actual verification, it will check one by one according to the actual characteristics of the checked data according to the rules of the rule set, and form the set of calibration results for each class of data.

TABLE I. CHECKOUT TABLE OF REPETITION RATE

Indicator	Check rule	Target	Data center	Original table	Upload table
Repetition rate	Repeat data account for the total amount of data	Checking field	MRID	SBID	SBID
		Data amount	1776	1799	1798
		Repeat data amount	76	6	42
		Repetition rate results	Repeated data amount/data amount=76/1776=4.28%	Repeated data amount/data amount=6/1799=0.33%	Repeated data amount/data amount=42/1798=0.23%

TABLE II. CHECKOUT RULES OF DATA QUALITY

Check dimension	Check scale	Check rule	Check indicator
Redundancy	Data set	Duplicate data contents are not allowed in different records in the data set, such as the business primary key or the data primary key duplication	Data redundancy rate
	Single data	Single data of the same record does not allow for fields that are not required by business or logical judgment	Field redundancy rate
Integrity	Data set	After removing redundant data, the difference of total amount of data set in the link of adjacent data flow should be within a reasonable range	Data loss rate
	Single data	Some fields are not allowed to exist empty values on the same record	Date integrity rate
Accuracy	Single data	A format that does not allow certain fields to exceed the value of the business requirements or not meet the requirements of the business on the same record	Data accuracy rate
Consistency	Single data	The content of a data recorded in a data set has a consistent relationship with the contents recorded by other data sets, such as values, logical computing relationships, etc.	Data consistency rate
Timeliness	Single data	The integration time of data records must be within the allowed time range	Data timeliness rate
Intelligence	Data set	The relationship between a data record of a dataset and the data records of other data sets	Date integration rate
	Single data	A data record in the original table can be associated with the rules in the data record in the upload table	Data conversion success rate

IV. CASE ANALYSIS

This paper choose transformer account data of power transmission and transformation from one province. According to the research findings, account data of power transmission and transformation comes from production and management system (PMS), it is transmitted through a series of links and uploaded to the head quarter of the power system at last. PMS system pushes the modified or the updated data into data center through batch task termly, the data key field is always the same. In addition, when the pushed data's key field is the same to the data in data center, it will update the data. The data center is associated with a number of data tables in accordance with certain view rules and extract data which the power system needs. The data is collected to the province side of power station original table. Data in original table will be converted to upload table data followed certain conversion rule. In the end, the power station will upload the data from upload table into the head quarter followed certain uploading rule and then be storage officially.

In this paper, the redundancy, integrity, intelligence, consistency, timeliness and accuracy of the data are checked by the data of the transformer ledger. Each check index has its specific check rules. Taking data repetition rate and data consistency rate as an example, the related fields and calculation process are as Table I and Table II.

Each check indicator is checked in accordance with the respective check rules. The checksum is summarized and the results are as Table III and IV.

We can draw some conclusions from Table III and IV:

- There is a lot of redundant data in the data center, which can easily affect the efficiency of data acquisition.
- The two links of the transformer account from the PMS to the data center and the upload table to the headquarters power system are the most likely to lose data.
- The data inconsistencies in the conversion process of the original table to the upload table of the transformer ledger are serious.
- In the process of integrating transformer accounts to headquarters, data acquisition and data conversion are most easily caused by untimely data.
- There are many data fields that do not meet the business requirements when the transformer table is designed on the provincial side data sheet.

Through the verification of the data quality of the transformer, the verification model made in this paper is reasonable. According to the results of data validation, we can get a detailed understanding of the data quality problems in the process of data transfer, and provide a reference for subsequent data quality governance.

V. CONCLUSION

The development of big data technology provides a basis for data quality evaluation. Based on big data and blood relationship theory, this paper establishes detailed checkup rules and evaluation system for power quality data. The research of the data quality evaluation system of the power quality on-line monitoring system is of great significance to improve the power quality effectively and improve the level of the power supply service.

This paper uses a theoretical data blood power quality data of online monitoring system of quality evaluation method based on the check, check with the specific business rules through the design, improve the data field automatically integrated ratio, the amount of data integrity, consistency and accuracy, reduce the data redundancy rate and deviation rate of power system data. Whether conform to relevant laws or whether there is abnormal situation, the power quality index early warning, to enhance the level of power supply reliability management, at the same time for the future power system data analysis and data mining provide good data base.

TABLE III. CHECKOUT TABLE OF CONSISTENCY RATE

Indicator	Check rule	Target	Data center-Original table		Original-Upload table		Upload table-Head quarter	
			Data center	Original table	Original table	Upload table	Upload table	Head quater
Data consistency rate	Data amount whose key field's content is the same account for the total amount of data	Matching field	MRID1	obj_id	obj_id	obj_id	obj_id	obj_id
		Checking field	OPERATION DATE	tyrq	tyrq	tyrq	tyrq	tyrq
		Matching data amount	1781		1759		1539	
		Data amount whose operation dates are the same	1760		1379		1518	
		Consistency rate results	Data amount whose operation dates are the same/Matching data amount=1760/1781=98.82%		Data amount whose operation dates are the same/Matching data amount =1760/1781=78.40%		Data amount whose operation dates are the same/Matching data amount =1760/1781=98.64%	

TABLE IV. SUMMARIZED RESULTS OF TRANSFORMER DATA QUALITY

Dimensions	Index	PMS	Data center	Original table	Upload table	Head quarter
Redundancy	**Data repeat rate**	——	4.28%	0.33%	0.23%	——
	Field redundancy rate	——	——	73.60%	39.74%	39.74%
Integrity	**Data loss rate**	6.69%	0%	2.06%	22.72%	
Intelligence	**Data automatically integrate rate**	94.41%	99.77%	95.77%	13.63%	
Consistency	**Data consistency rate**	——	98.82%	78.40%	98.64%	
Timeliness	**Data timeliness rate**	——	100%	31.37%	23.08%	——
Accuracy	**Data accuracy rate**	——	——	69.92%	88.98%	——

REFERENCES

[1] Zong Wei and Wu Feng, "Chanllenge of Data Quality in the Big Data Age," Journal of Xi'an Jiaotong University(Social Sciences), vol. 33, Sep. 2013, pp. 38-43

[2] Caballero I, Serrano M and Piattini M, "A Data Quality in Use Model for Big Data,"Springer International Publishing, Vol. 63, 2014, pp. 123-130

[3] Pipino L L, Yang W L and Wang R Y, "Data Quality Assessment," ACM, 2002

[4] Rakhshani, E. Sariri I, Rouzbehi K, "Application of data mining on fault detection and prediction in Boiler of power plant using artificial neural network," International Conference on Power Engineering Energy and Electrical Drives, 2009, pp. 473-478

[5] Yi-Ting Huang, Fan-Tien Cheng. "Automatic Data Quality Evaluation for the AVM System," IEEE Transactions on Semiconductor Manufacturing, Vol. 24, 2011, pp. 45-54

[6] Qin Xuan, "Data Quality Evaluation and Anomaly Detection Research based on the Statistical Data of Power System," Changsha University of Science & Technology, April. 2013.

[7] Zhou Hui, Yang Honggeng and Wu Chuanlai, "A Power Quality Comprehensive Evaluation Method based on gre clustering," Power System Protection and Control, Vol. 40, Aug. 2012, pp.70-75

[8] Zhang Xiaohan, Lyu Ying, Bai Jingqiang, Tan Jun and Lu Jingjing, "Improvement of Power Supply Reliability Based on Power Quality Onlne Monitoring", Distribution & utilization, 2015, pp:10-16

[9] Wang Linxin, Mu Junqiang and Ju Huifang, "Study and Application of On-line Power Quality Monitoring and Measuring System," Electric Power Information and Communication Technology, Vol. 13, Feb. 2015, pp.132-136

[10] Chiticariu L, Tan W C and Vijayvargiya G. DBNotes, "a post-it system for relational databases based on provenance," ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, Usa, June. DBLP, 2005, pp. 942-944

[11] Fan H, "Tracing Data Lineage Using Automed Schema Transformation Pathways," Advances in Databases. Springer Berlin Heidelberg, 2002, pp. 50-53