A Big Data Framework for Electric Power Data Quality Assessment

He Liu[1], Fupeng Huang[1]
[1]Advanced Computing and Big Data Technology Laboratory of SGCC,
Global Energy Interconnection Research Institute,
Beijing, China
liuhe, huangfupeng@geiri.sgcc.com.cn

Han Li[2,3], Weiwei Liu[1], Tongxun Wang[4]
[2]Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing, China
[3]College of Computer Science, North China University of Technology, Beijing, China
[4]State Key Laboratory of Advanced Power Transmission Technology,
Global Energy Interconnection Research Institute,
Beijing, China
lihan@ncut.edu.cn, liuweiwei,
txwang@geiri.sgcc.com.cn

***Abstract*—Since a low-quality data may influence the effectiveness and reliability of applications, data quality is required to be guaranteed. Data quality assessment is considered as the foundation of the promotion of data quality, so it is essential to access the data quality before any other data related activities. In the electric power industry, more and more electric power data is continuously accumulated, and many electric power applications have been developed based on these data. In China, the power grid has many special characteristic, traditional big data assessment frameworks cannot be directly applied. Therefore, a big data framework for electric power data quality assessment is proposed. Based on big data techniques, the framework can accumulate both the real-time data and the history data, provide an integrated computation environment for electric power big data assessment, and support the storage of different types of data.**

Keywords-data quality; electric power data; data quality assessment; big data; framework

I. INTRODUCTION

With the development of IoT technology, a large amount of data is produced. Lots of statistics show that the total amount of data generated in China exceeded 0.8ZB in 2013, which is two times the amount of data in 2012. By the end of 2020, it is estimated that the total amount of data generated in China will be ten times the amount of data in 2013, that is, the big data era has come.

In the age of big data, data has become one of the most important factors, and lots of applications are provided on the basis of these data. In recent years, more and more researchers and corporations begin to attach great importance to the big data techniques, such as BAT and Ali. That is, data has turned into the core assets, and the value of these assets has attracted numerous attentions [1].

Data quality is a nebulous term. In this paper, data quality is defined as the totality of features and characteristics of data that bears on its ability to satisfy a given purpose [2]. In the past few years, the number of data quality problem increases year by year, and the data quality problem raises the organization cost. Studies have shown that the cost caused by the low-level data is about 8%~12% of the organization income, and the highest proportion even reached 40%~60%. As a result, data quality is considered as the foundation of all data related activities.

Due to the universality and inevitability of data quality problem, how to provide high-quality data has become a new research hot spot in the field of big data. Since the promotion of data quality is not an arbitrary thing, the quality of data should be properly assessed before data cleaning. Therefore, this paper takes the electric power big data acquired in the power grid of China as the research object, and an electric power data quality assessment oriented big data framework is proposed.

II. RELATED WORK

In the big data era, if the data quality is not well managed, it is impossible to achieve effective data processing and data mining. At present, more attentions have been paid to how to mine valuable information from the big data, while the research on the quality of big data is not enough.

In recent years, many data quality assessment techniques are proposed in various fields. In 2010, a framework to assess data quality in university web portals is discussed [3]. In the framework, essential data quality dimensions are captured in four categories. But it is only a generic framework, and cannot be applied in the electric power industry. In 2011, the information content of dimensions and indicators of a data quality assessment framework is investigated by applying information entropy theory [4], and some suggestions are brought forward on improving the data quality assessment framework. However, the study mainly concentrates on the statistical data quality. In 2016, a framework to assess healthcare data quality is proposed to avoid invalid conclusions and misinformed management decisions in healthcare applications [5], and a range of criteria are defined. However, the framework mainly focuses on the criteria of data assessment. In the same year, a conceptual methodology for assessing linked datasets is described, and a framework called Luzzu is proposed for

978-1-5386-4806-3/17 $31.00 © 2017 IEEE
DOI 10.1109/WISA.2017.29

linked data quality assessment [6]. By analyzing a number of statistical datasets against a variety of metrics, Luzzu is proved to be able to improve the quality of linked data, but it is not suitable for electric power data assessment. Although the above researches have pay attention to the framework of data assessment, they are not suitable for assessing the quality of big data.

Nowadays, researchers have begun to pay their attentions to big data quality assessment techniques. In 2015, a framework to assess the quality of big data is designed based on the decision tree and the multidimensional model [7]. The paper mainly concentrates on the dimensions of big data assessment, but does not consider the characteristics of different fields. In 2016, the "3as data quality-in-use model" is proposed to achieve big data quality assessment [8]. In the model, three data quality characteristics are applied for assessing the quality of big data. However, the framework is more suitable for quality-in-use of big datasets. In 2016, an efficient data quality evaluation scheme by applying sampling strategies on big data is discussed [9]. Although the schema is proved to be feasible, the evaluation object is not the entire dataset. Meanwhile, a scalable assessment approach for big data quality evaluation is proposed [10], and an initial prototype to investigate scalability in a multi-node test environment is accomplished using big data technologies. However, the proposed approach is particular emphasis on the big data in smart ecosystems.

Except for data assessment techniques, data acquisition techniques and data storage techniques are also required when the electric power big data is assessed. Recently, although many distributed techniques are proposed for massive data collection and storage [11], they are not able be directly applied for electric power big data.

To sum up, this paper concentrates on the big data framework for the electric power data assessment.

III. The Proposed Framework

By adequately analyzing the characteristics of the electric power big data in the power grid of China, a big data framework for electric power data quality assessment is proposed.

A. Characteristics of Electric Power Big Data

Due to the special characteristics in the aspect of volume, variety, velocity, variability and veracity, the processing and storage issues of big data always exceed the capability of traditional information technology. The primary characteristics of electric power big data in the power grid of China are listed as follows.

1) Mutilayer sturcture: There are multiple layers in the power grid, such as the headquarters, the provincial power grid, the prefectural power grid, the municipal power grid and so on. From the perspective of the headquarters, the power grid is composed of two levels which are the headquarters and the provincial power grid.

2) Large scale: Sine data are continuously generated from the massive electric power sensors which are widely deployed in the smart gird, the scale of these data is sharply increases. For example, the data capacity of harmonic detection data from a province in China will reach 3T, when the harmonic data are produced by 1000 monitoring sites, each monitoring site contains 2000 indicators.

3) Mutiple data types: There are primarily three kinds of electric power data, including the fundamental data, the history data and the real-time data. The fundamental data are used to describe sensor devices, indicators and so on. The history data include the waveform data and files which contain the historical running statuses of the sensor devices. The real-time data refer to the data continuously generated by the sensor devices.

4) Mutisource data: Different types of data are produced by different ways. The fundamental data is manual entered. The real-time data is continuously generated by sensor devices. The history data is accumulated from the sensor devices.

5) Information island: Since sensor devices are deployed in every provicial power grid, data are not integrated, and in-depth analysis is unable to be accomplished from the overall perspective.

6) Different performance requirements: According to the demands of different electric power applications, the time to achive a data processing varies from minute to hour.

Considering the above characteristics, a framework for electric power big data is designed instead of the standard big data framework, and a data assessment module is involved in the framework to promoting the quality of electric power data.

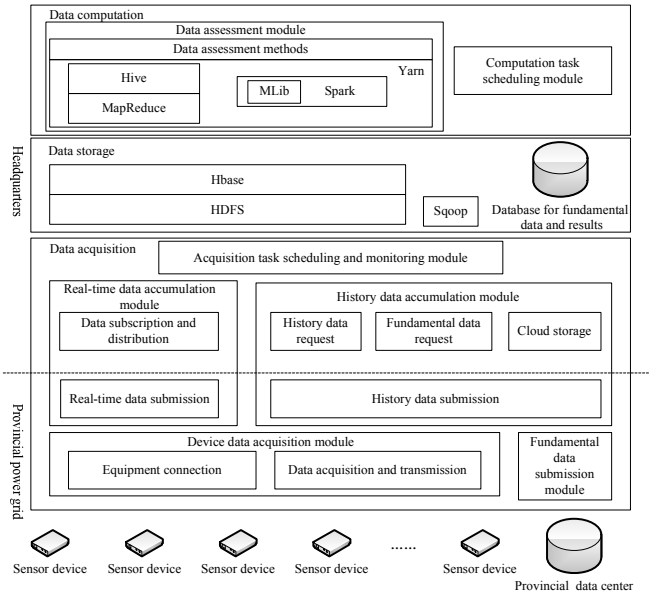

Figure 1. The proposed framework

B. Framework

Considering the above characteristics, an electric power data quality assessment oriented big data framework is designed in Figure 1. In the aspect of the function, it comprises three parts which are data acquisition, data storage and data computation. In the aspect of the organizational

structure, it is divided into two parts, including the headquarters and the provincial power grid.

1) Data Acquisition: Data acquisition is an essential part in nearly every big data system. As illustrated in Figure 1, data acquisiton comprises four main modules. The details are dipiected as follows.

a) Device data acquisition module: This module locates in the provicial power grid. It is used to provide data by collecting running data from plenty of sensor devices.

b) Real-time data accumulation module: This module takes charge of gathering the real-time data from the provincial power grid. The process of real-time data accumulation is shown in Figure 2. The headquarters firstly obtains the configurations and topics of Kafka from the provincial power grid. Secondly, data is received by multiple threads. Finally, the real-time data is saved in HBase, and the accumulating logs are recorded.

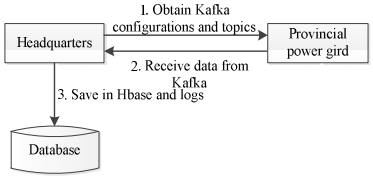

Figure 2. The process of real-time data accumulation

c) History data accumulation module: This module is responsible for obtainng history data from the provincial power girds. The process of real-time data accumulation is shown in Figure 3. A socket connection is firstly established between the headquarters and the provincial power grid. Secondly, the headquarters obtains the account and password of the ftp server deployed in the provincial power grid. Thirdly, the file names of history data are obtained. Fourthly, the zip files which contain the history data is downloaded and unzipped. Finally, the history data is saved in HBase and HDFS, and corresponding logs are recorded.

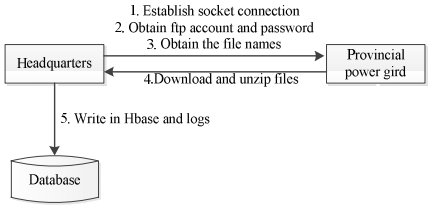

Figure 3. The process of history data accumulation

d) Acquisition task scheduling and monitoring module: Since the scale and frequency of history data are not the same, this module is used to schedule tasks on the basis of the monitoring data.

2) Data Storage: To store different types of data, an integrated storage environment is designed, which includes a relational database Oracle, a NoSQL database HBase and a distributed file system HDFS. Oracle is used to save structured data including both the fundamental data and the results of assessment, HDFS is applied to receive the history data, and HBASE is used to save both the real-time data and the information extracted form the history data.

3) Data Computation: Data computation is composed of two main modules whose details are depicted as follows.

a) Data Assessment module: This module is in charge of checking and evaluating the quality of electric power big data using various data quality assessment methods.

As shown in Figure 1, all of the data quality assessment methods are supported by a big data platform, which involves various big data techniques, such as Sqoop, Hive, Hbase, HDFS, MapReduce (MR) and Spark.

Data quality is a multi-dimensional concept. Subjective data quality assessment and objective data quality assessment are two primary types. Subjective data quality assessment is achieved based on the experiences of stakeholders. Objective data quality assessment mainly depends on the states of data. Considering both the objective data quality assessments and the subjective data quality assessment, the process of data quality assessment applied in the framework is illustrated in Figure 4.

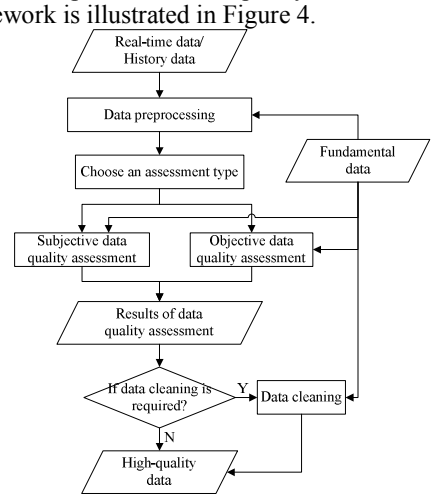

Figure 4. The process of data quality assessment

As shown in Figure 4, the input includes the real-time data, the history data and the fundamental data, while the output refers to the results of data quality assessment and the high-quality data. Since the structure of data may not suitable for the subsequent assessment, preprocessing is firstly applied to adjust the structure of either the real-time data or the history data. For example, redundant blank lines and blanks are removed. Secondly, the type of data quality assessment is determined. For data whose quality is not able to be assessed by objective data quality assessment methods, subjective data quality assessment is adopted. Thirdly, the results of data quality assessment are evaluated so as to make sure whether the input data need to be cleaned. If the quality of the input data is low, data cleaning methods are applied, such as the threshold based outlier detection method and the k-means based outlier detection method.

b) Assessment task scheduling module: Due to the large number of assessment tasks and the constraints

between these tasks, it is necessary to schedule these assessment tasks, such as scheduling the tasks according to the time constraint.

C. Implementation

As shown Figure 5, in The proposed framework is implemented to support the electric power big data assessment in the power grid of China.

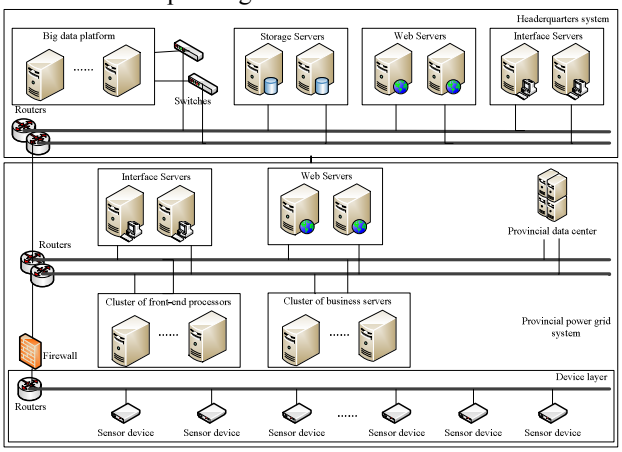

Figure 5. The implementation of the proposed framework

1) Headquarters system: There are four primary parts. Interface server is applied to communicate with the business servers in the provincial power grid system. Web server is used to support different applications. Storage server contains the relational data for the results of assessment and the fundamental data. Big data platform is used to save the real-time data and the history data, and to support the data quality assessment methods.

2) Provincial power grid system: There are five primary parts. Interface server is responsible for achieving the communication between the headquarters system and the provincial power grid system. Web server is deployed to support local electric power related applications. Provincial data center takes charge of saving the fundamental data and the history data. Cluster of business servers is a set of servers which are used to obtain, analyze and cache data. Cluster of front-end processors is a set of front-end processors. A front-end processor is a device used to preprocess the data which is collected from the sensor devices. Moreover, the hot standby technique which is able to avoid single point of failure.

IV. Conclusion

In the big data era, the majority of applications are driven by data. If the data quality can be promoted, the effectiveness and reliability of the applications are able to be improved as well. In the power grid of China, the electric power big data has a number of special characteristics in the aspect of the scale, the structure, the format and the assessing requirements. In order to adapt to these characteristics, an electric power data assessment oriented big data framework is proposed. Based on a variety of big data techniques, the proposed framework is able to accumulate and store the real-time data, the history data and the fundamental data, and support the high-performance data quality assessment. The implementaion of the proposed framwork proves its feasibility, and the proposed framework is able to provides a valuable framework for other big data applications with similar characteristics. Besides, involving more data assemment methods and guarteeing the data quality during data accumuation are the future work.

Acknowledgment

This work is supported by State Grid Company Research Project under grant SGRIJSKJ[2015] 1029.

References

[1] R. Singh, K. Singh, “A descriptive classification of causes of data quality problems in data warehousing.” International Journal of Computer Science Issues, vol. 7, no. 3, pp. 41-50, 2010.

[2] J. H. Swift, IOCCP Report on Reference-Quality Water Sample Data: Notes on Acquisition, Record Keeping, and Evaluation. 2010.

[3] M. A. Mary, A Framework to Assess Data Quality in University Web Portals, Master Thesis, India: Delft University of Technology, 2010.

[4] X. Wang, Y Tetsuya, F. Kong, A. Yuki H. Yoichiro, “Data quality assessment framework and its information content analysis,” Journal of Business Economics, vol. 3, no. 4, pp.322-332, 2011.

[5] W. Warwick, S. Johnson, J. Bond, G. Fletcher, P. Kanellakis, “A framework to assess healthcare data quality,” The European Journal of Social and Behavioural Sciences, vol. 13, no. 2, pp. 1730-1735, 2015.

[6] J. Debattista, S. Auer, C. Lange, “Luzzu-A methodology and framework for linked data quality assessment,” Journal of Data and Information Quality, vol. 8, no. 1, pp. 1-4, 2016.

[7] D. V. Subramanian, K. Pradheepkumar K. Dhinakaran, Duraimurugan, “Catur approach to assess the quality of big data using decision tree and multidimensional model,” Australian Journal of Basic and Applied Sciences, vol. 9, no. 23, pp. 503-508, 2015.

[8] M. Jorge, C. Ismael, R. Bibiano, S. Manuel, P. Mario, “A data quality in use model for big data,” Future Generation Computer Systems, vol. 63, no. C, pp. 123-130, 2016.

[9] T. Lkbal, K. H. T. Ei, S. M. Adel, D. Rachida, B. Chafik, “Big data quality: A quality dimensions evaluation,” Proceedings of the 13th IEEE International Conference on Ubiquitous Intelligence and Computing, pp. 759-765, July 2016.

[10] M, Klas, W. Putz, T, Lutz, “Quality evaluation for big data: a scalable assessment approach and first evaluation results,” Proceedings of the International Conference on Software Process and Product Measurement, pp. 115-124, October 2016.

[11] L. G. Abdulkhader, V. Sanja, J. Valentina, “Big data and quality: A literature review,” Proceedings of the 24th Telecommunications Forum, pp.1-4, November 2016 .