Analytic Model and Assessment Framework for Data Quality Evaluation in State Grid

Zhe Li$^{1,a}$, Sai Wu$^{1,b}$, Hongwei Zhou$^{2,c}$, Sheng Zou$^{2,d}$ and Tingting Dong$^{1,e}$

$^1$China Electric Power Research Institute, Beijing, 100192, China
$^2$State Grid Jiangsu Economic Research Institute, Nanjing, 210000, China
$^a$Email: lizhe@epri.sgcc.com.cn
$^b$Email: wusai@epri.sgcc.com.cn
$^c$Email: zhouhongwei@js.sgcc.cn
$^d$Email: zous@js.sgcc.com.cn
$^e$Email: xintong-dongting@epri.sgcc.com.cn

Abstract. With the construction of information infrastructure and the smart grid development, big data platform become an important component to support the construction of smart grid, as a consequence, the quality of the data contain in the big data platform become an important part of the process of big data platform construction. In this paper, a discussion about the current data quality analysis model was presented and also the typical data quality problem appears in the process of information infrastructure construction of the State Grid Corporation is analyzed. On the basis of this, the data quality assessment for the information system in the whole life cycle of the State Grid Corporation is designed. As a result, data quality can be checked by the analysis model in order to ensure the data quality level in the whole life cycle of the information system, finally a case was carried out for the preliminary analysis by using the mentioned framework, the result show that this framework can be effective in embedded into the workflow and can be helpful for the data quality assessment.

1. Introduction
In the recent years, with the wide use of information system to support smart grid construction, a large volume of data has been produced and stored in different databases. As an important application, big data platform have received much attention from both research and industry area, Based on an IDC report prediction, the global data volume will grow exponentially from 4.4 zettabyte to 44 zettabytes between 2013 and 2020$^{[1]}$. By 2025, IDC predicts there will be 163 zettabytes of data$^{[2]}$. One question for large enterprises is determining who should own big-data initiatives that affect the entire organization$^{[3]}$.

With the construction of the information system, the data become more and more valuable for the upgrade of the service, and also data center has become one of the most important fundamental facility, which lots of applications extract the data from. Moreover, after years construction of the data center, data itself has change to a more important asset not only for the construction of the new application, but also for the value of the essential customer.

Data Quality problem in the real world can be divided into two categories: grammatical errors and semantic errors. Grammatical errors refer to the violation between the constraint and the value while the semantic error refers to the inconsistency between the values in the data and the values represented. Traditional database systems usually focus only on the amount of data, supporting the creation, maintenance and use of large amounts of data, but if the stored data itself has problems, that is, the
queried data is dirty, that is, the data does not represent the real world entities, such a database system cannot find the correct answer[4]. When dirty data exists in the database, no matter how powerful the query and creation ability of the database is, the results of data query do not conform to the entities in the real world. Data quality is part of the data that the user expects to execute, or the data value that the user expects to get. From this definition, we can see that the concept of data quality is relative. Just as different users have different understandings of the data itself, the requirements for data quality in data warehouse will of course vary with different observers or users.

Data quality is defined as if the data can be mapped into the real world correctly, and also can reflect the real world in given constant, which means the data is not only a value stored in the database, but also need to represent the attribute which belong the real thing in the real world and also in the real time. Data quality exists in the operation of data by other modules of data warehouse, which reflects the inherent attributes of data in data physical layer. It can be divided into data integrity, first-hand data reliability, data accuracy, data consistency and data uniqueness. From a quantitative point of view, integrity can be considered as the number of illegal null value records in data warehouse; credibility can be measured by the number of records carrying default values in data supermarkets. However, there are two main difficulties exist in the construction of the big data center, one is how to design a rule to make sure the data quality in a certain level in the design stage. Second is how to make sure the data quality can be maintained during the operation stage.

The important factor which can be used to settle the difficult is the how to model analysis the data quality towards given data set and how to find the problem which may affect the data quality. In this paper, we designed the data quality analytic model and give a case we use this model to assess the data quality of the given data set.

The rest of the paper is organized as follows. Section2 describe the related work while Section 3 will define the methodology. However, the experiment and result analysis is elaborated in Section 4. At the end, Section 5 summarizes our conclusions and highlights future direction.

2. Related Work

Different entities in the real world have different representations and also can be varies from different system applications, however, it may be correct for given attribute in specific aspect while may not correct to integrate different attribute from all the data source. Lots of work have been work on the different aspects of the data quality. Sort and merge algorithm is a similar entity recognition method which can be used to sort the attributes that does not depend on domain-specific knowledge[5]. Fellegi-Sunter model can be used to similar record recognition using context for the recognition of uncertain redundancy information[6].

To assess the data quality, a framework to model essential data quality dimensions are captured in four categories[7], after then entropy theory was applied in the data quality assessment framework while also give out some advice toward to the improvement of data quality[8]. To assess the linked data the against a set of metrics, Luzzu is proved to be able to improve the quality of linked data[9]. Some application of the assessment framework to assess data quality has also been proposed in different application, such as in healthcare applications[10]. Although the researches mentioned above have great contribution to the framework of data assessment, they are not applicable for assessing the quality in big data scenario.

To capture the big data quality feature, a framework is designed based on the decision tree and the multidimensional model[11]. To simplify the data quality assessment process, an data quality evaluation scheme by applying sampling strategies on big data and also a scalable assessment approach was designed and an initial prototype to investigate scalability in a multi node test environment is accomplished using big data technologies[12, 13].

Although there are many methods involved in the above research, in the practical application scenarios, how to design and implement the data quality assessment framework for specific domain have been paid little attention in the research area. Therefore, this paper gives a data quality assessment framework for the power field, while it also gives a case study for the application of the data quality assessment framework in specific domain.
3. Methodology

Data quality evaluation system consists of data quality objectives, data quality dimensions and data quality constraints. The goal of data quality is a judgment framework under the semantic condition of large data analysis. Data quality dimension is used to describe data quality from different perspectives. Data quality checking is carried out after the completion of table data development, which mainly includes four aspects: duplicate value checking, missing value checking, data skew problem and outlier checking. The missing values are mainly the missing information of some fields in some records of the index dataset. Data skew problem: The distribution of field values is mainly concentrated in a specific category or a specific interval.

3.1. GQM Based Evaluation Model

Data quality model is the combination of data quality evaluation object, target and data quality evaluation system. It includes the key steps of identifying evaluation target and requirement, forming data set to be evaluated, determining evaluation index and weight, data quality diagnosis and quantitative calculation, and forming evaluation results.

In this paper, GQM model is used to evaluate data quality. It is a goal-oriented, top-down definition method of software metrics[14]. For data quality measurement, the model first determines the quality target of data instances, secondly tracks the target to the problem. Finally, these problems need to be able to define the goal operatively, and give a framework to explain the goal and problem. Because each data instance needs to achieve a series of goals, and each goal needs to answer a series of questions.

![Figure 1. GQM Based Evaluation Model](image)

For each question, we can find a corresponding complete and quantifiable solution, through which we can find the data quality measurement of the target. Therefore, GQM model can be used to summarize and decompose data quality objectives or data instance objectives to form measurement indicators, and then extract the values that can be used to measure from these indicators, so as to achieve the purpose of prediction, process control and quantitative evaluation of data quality.

3.2. Data Quality Diagnosis and Quantitative Computation

According to the verification rules, the data that violates the relevant rules are found, and the problem diagnosis of the evaluated data sets is realized. At the same time, it is advisable to use the quantitative method to carry out the quantitative calculation of each index subitem. Quantitative values of evaluation indicators are weighted by the results of quantitative calculation of indicators subitems. 

\[
O = \sum_{i=1}^{k} w_i I_i
\]

in which \(O\) is the quantitative results of evaluation indicators, \(k\) is the number of the subitem of indicators, \(I_i\) is the quantitative results of subitem \(i\)

\[
\sum_{i=1}^{k} w_i < 1
\]
3.3. Data Quality Assessment Process

In the process of data quality assessment, it is usually necessary to quantify the quality of the data set assessed around the data quality dimension. Figure 2 shows the relationship among the data quality dimension, evaluation indicators and quantitative description.

![Figure 2. Data quality assessment process](image)

Data quality analysis model needs to extract different data sources for a given system, define different dimensions of data quality analysis, such as correctness, completeness, consistency, uniqueness, accuracy and validity, and define the weight of data quality evaluation defined by different data sources and data quality analysis dimensions, as well as the number of the whole system by the quality itself. Comprehensive Weighted Average Based on Quality

4. Experiment and Result Analysis

4.1. Description of Case

The Shandong company of SGCC carries a large number of information systems, a total of 102 sets (classes), 102 sets (classes) of information system names and their deployment methods are detailed in Table 4-1, including 63 first-level systems, 37 second-level systems and 2 external systems. In the list of self-built systems of State Grid Shandong Company of SGCC.

Within the scope of the company, the main types of system are: 26 business operations applications, 56 management and control applications, 3 assistant decision-making and 17 support applications. However, the company has not yet managed these data as a whole, which further results in a lot of redundant data and temporary process data in the company's data center. In terms of data acquisition, there are still some deficiencies in business integration, data acquisition standards, equipment utilization, data unification, classification and distribution, etc.

![Figure 3. Evaluation process in Case Study](image)

As it is shown in Figure 3, to evaluate the data quality of the company, we first pre-treat the data with extract, parse and collate the data sources according to the method proposed in section 3. On this basis, we select the corresponding data quality evaluation methods from different dimensions of data quality requirements, different attributes of data and data characteristics under large data scenario, and carry out different data types. Measure, and ultimately integrate all the evaluation results to form the overall evaluation.
4.2. Application of GQM Based Assessment

In order to integrate the methods proposed in section 3, we embedded the data quality evaluation

The four steps illustrated in Figure 4 is

a) Establishment of assessment objectives and requirements

The objectives and requirements of data quality assessment are determined by the management and implementation of data quality assessment tasks.

b) Determine the data set to be evaluated

According to the evaluation objectives and requirements, the scope of the data to be assessed is determined, the data set to be assessed is formed, and auxiliary materials such as data dictionaries and business descriptions are prepared.

c) Determine the evaluation index and weight

According to the data quality evaluation system, the data quality dimensions involved in the evaluation process are determined, and the evaluation indicators and sub-indicators are formed. Assessment indicators include one or more subitems of indicators and the weight distribution among sub-items of indicators, which are formulated by the implementation of the evaluation. The generation of evaluation results of indicator subitems, and the execution of one or more verification rules. As a result, the verification rules should be formulated according to the constraints of the data quality assessment system.

d) Large data sampling

The data included in the evaluation method are sampled according to the quality objectives of large data and the data set to be evaluated, and the sampling distribution of the preliminary sampling evaluation estimators is carried out.

4.3. Result Analysis

According to our Assessment, 56109 data items was checked while 2664 items need to be improved, of which 2496 items have been revised, 143 items have been added and 25 items have been deleted, resulting in 56084 items of Shandong Company's data inventory.

| Domain                      | Delete(D) | Modify(M) | ADDX(A) | Total |
|-----------------------------|-----------|-----------|---------|-------|
| Technology Management       | 0         | 811       | 0       | 811   |
| Information and communication| 0      | 732       | 0       | 732   |
| Support                     | 0         | 545       | 0       | 545   |
| Regulatory restrictions     | 0         | 216       | 0       | 216   |
| Outreach                    | 0         | 102       | 0       | 102   |
| Operation                   | 1         | 23        | 27      | 101   |
| Marketing                   | 22        | 33        | 45      | 100   |
| Finance                     | 2         | 7         | 21      | 30    |
| Safety supervision          | 0         | 24        | 0       | 24    |
| Union                       | 0         | 2         | 0       | 2     |
| Supplies                    | 0         | 1         | 0       | 1     |
| Total                       | 25        | 2496      | 143     | 2664  |

The overall difference mainly concentrates on finance, operation inspection and marketing. These three systems need to add 143 data items, modify 63 data items and delete 25 data items. Specifically, first, 21 new data items are added in finance, 7 data records are modified and 2 data items are deleted; second, 77 new data items are added in operation inspection, including 25 self-designed distribution automation systems, 23 data records are modified and 1 data item is deleted; third, 45 new data items
are added in marketing system, 33 data records are modified and 22 data items are deleted. Among the modification records, there are 18 modifications of "whether or not first-hand data", 22 modifications of structured data, 8 modifications of table names, 3 modifications of data sources and 12 modifications of data item names. revised 2433 items of data items in rest system, all of which suggested that unstructured data items should be converted to structured data items.

5. Conclusion
In this paper, we designed a framework for data quality analysis aiming at all kinds of data quality problems existing in the SGCC information system based on GQM and data quality evaluation method, to validate the validation of the framework we verifies it in Shandong Company. The results show that the method can be well embedded in the company's business process, and can provide strong support for the company data quality check and inspect. In the process of research, we also find that there are important differences between the inspection and evaluation of data quality between structured data and unstructured data. Therefore, how to define the data quality characteristics of these two kinds of data and the quality feature under this two different data can be the further work.

6. Acknowledgments
This work was supported by State Grid Technology Project Grant(B3441618K002).

7. References
[1] Hajirahimova, M.S. and A.S. Aliyeva, About big data measurement methodologies and indicators. International Journal of Modern Education and Computer Science, 2017. 9(10): p. 1.
[2] Reinsel, D., J. Gantz, and J. Rydning, Data age 2025: The evolution of data to life-critical. Don’t Focus on Big Data, 2017.
[3] Newswire, F., Mastering big data: CFO strategies to transform insight into opportunity. 2012.
[4] Pipino, L.L., Y.W. Lee, and R.Y. Wang, Data quality assessment. Communications of the ACM, 2002. 45(4): p. 211-218.
[5] Dittrich, J.-P., et al. Progressive merge join: A generic and non-blocking sort-based join algorithm. in Proceedings of the 28th international conference on Very Large Data Bases. 2002. VLDB Endowment.
[6] Winkler, W.E., Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. 2000: US Bureau of the Census Washington, DC.
[7] Arul Mary, M., A Framework to assess Data Quality in university web portals. 2010.
[8] Stvilia, B., et al., A framework for information quality assessment. Journal of the American society for information science and technology, 2007. 58(12): p. 1720-1733.
[9] Debatista, J., S. Auer, and C. Lange. Luzzu--A Framework for Linked Data Quality Assessment. in 2016 IEEE Tenth International Conference on Semantic Computing (ICSC). 2016. IEEE.
[10] Kerr, K., T. Norris, and R. Stockdale, Data quality information and decision making: a healthcare case study. ACIS 2007 Proceedings, 2007: p. 98.
[11] Behrisch, M., et al. Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. in 2014 IEEE Conference on Visual Analytics Science and Technology (VAST). 2014. IEEE.
[12] Batini, C., et al., Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 2009. 41(3): p. 16.
[13] Kläs, M., W. Putz, and T. Lutz. Quality evaluation for big data: a scalable assessment approach and first evaluation results. in 2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA). 2016. IEEE.
[14] Hussain, A. and E. Ferneley. Usability metric for mobile application: a goal question metric (GQM) approach. in Proceedings of the 10th international Conference on information integration and Web-Based Applications & Services. 2008. ACM.