

PROCEEDINGS PAPER

The Challenges of Data Quality and Data Quality Assessment in the Big Data Era

Li Cai^{1,3} and Yangyong Zhu²¹ School of Computer and Science, Fudan University, No. 220, Han Dan Road, Shanghai, China
lcai@fudan.edu.cn² Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China
yyzhu@fudan.edu.cn³ School of Software, Yunnan University, No. 2 North Road of Cui Hu, Kunming, China

High-quality data are the precondition for analyzing and using big data and for guaranteeing the value of the data. Currently, comprehensive analysis and research of quality standards and quality assessment methods for big data are lacking. First, this paper summarizes reviews of data quality research. Second, this paper analyzes the data characteristics of the big data environment, presents quality challenges faced by big data, and formulates a hierarchical data quality framework from the perspective of data users. This framework consists of big data quality dimensions, quality characteristics, and quality indexes. Finally, on the basis of this framework, this paper constructs a dynamic assessment process for data quality. This process has good expansibility and adaptability and can meet the needs of big data quality assessment. The research results enrich the theoretical scope of big data and lay a solid foundation for the future by establishing an assessment model and studying evaluation algorithms.

Keywords: Big data; Data quality; Quality assessment; Data science

1 Introduction

Many significant technological changes have occurred in the information technology industry since the beginning of the 21st century, such as cloud computing, the Internet of Things, and social networking. The development of these technologies has made the amount of data increase continuously and accumulate at an unprecedented speed. All the above mentioned technologies announce the coming of big data (Meng & Ci, 2013). Currently, the amount of global data is growing exponentially. The data unit is no longer the GB and TB, but the PB ($1\text{PB} = 2^{10}\text{TB}$), EB ($1\text{EB} = 2^{10}\text{PB}$), and ZB ($1\text{ZB} = 2^{10}\text{EB}$). According to IDC's "Digital Universe" forecasts (Gantz & Reinsel, 2012), 40 ZB of data will be generated by 2020.

The emergence of an era of big data attracts the attention of industry, academics, and government. For example, in 2012, the US government invested \$200 million to start the "Big Data Research and Development Initiative" (Li & Chen, 2012). *Nature* launched a special issue on big data (Nature, 2008). *Science* also published a special issue "Dealing with Data" (Science, 2011), which illustrated the importance of big data for scientific research. In addition, the development and utilization of big data have been spread widely in the medical field, retail, finance, manufacturing, logistics, telecommunications, and other industries and have generated great social value and industrial potential (Feng, Z. Y., Guo, X. H., Zeng, D. J., et al., 2013).

By rapidly acquiring and analyzing big data from various sources and with various uses, researchers and decision-makers have gradually realized that this massive amount of information has benefits for understanding customer needs, improving service quality, and predicting and preventing risks. However, the use and analysis of big data must be based on accurate and high-quality data, which is a necessary condition for generating value from big data. Therefore, we analyzed the challenges faced by big data and proposed a quality assessment framework and assessment process for it.

2 Literature Review on Data Quality

In the 1950s, researchers began to study quality issues, especially for the quality of products, and a series of definitions, for example, quality is “the degree to which a set of inherent characteristics fulfill the requirements” (General Administration of Quality Supervision, 2008); “fitness for use” (Wang & Strong, 1996); “conformance to requirements” (Crosby, 1988) were published. Later, with the rapid development of information technology, research turned to the study of the data quality.

Research on data quality started abroad in the 1990s, and many scholars proposed different definitions of data quality and division methods of quality dimensions. The Total Data Quality Management group of MIT University led by Professor Richard Y. Wang has done in-depth research in the data quality area. They defined “data quality” as “fitness for use” (Wang & Strong, 1996) and proposed that data quality judgment depends on data consumers. At the same time, they defined a “data quality dimension” as a set of data quality attributes that represent a single aspect or construct of data quality. They used a two-stage survey to identify four categories containing fifteen data quality dimensions.

Some literature regarded web data as research objects and proposed individual data quality standards and quality measures. Alexander and Tate (1999) described six evaluation criteria - authority, accuracy, objectivity, currency, coverage/intended audience, and interaction/transaction features for web data. Katerattanakul and Siau (1999) developed four categories for the information quality of an individual website and a questionnaire to test the importance of each of these newly developed information quality categories and how web users determine the information quality of individual sites. For information retrieval, Gauch (2000) proposed six quality metrics, including currency, availability, information-to-noise ratio, authority, popularity, and cohesiveness, to investigate.

From the perspective of society and culture, Shanks and Corbitt (1999) studied data quality and set up an emiotic-based framework for data quality with 4 levels and a total of 11 quality dimensions. Knight and Burn (2005) summarized the most common dimensions and the frequency with which they are included in the different data quality/information quality frameworks. Then they presented the IQIP (Identify, Quantify, Implement, and Perfect) model as an approach to managing the choice and implementation of quality related algorithms of an internet crawling search engine.

According to the U.S. National Institute of Statistical Sciences (NISS) (2001), the principles of data quality are: 1. data are a product, with customers, to whom they have both cost and value; 2. as a product, data have quality, resulting from the process by which data are generated; 3. data quality depends on multiple factors, including (at least) the purpose for which the data are used, the user, the time, etc.

Research in China on data quality began later than research abroad. The 63rd Research Institute of the PLA General Staff Headquarters created a data quality research group in 2008. They discussed basic problems with data quality such as definition, error sources, improving approaches, etc. (Cao, Diao, Wang, et al., 2010). In 2011, Xi'an Jiaotong University set up a research group of information quality that analyzed the challenges and importance of assuring the quality of big data and response measures in the aspects of process, technology, and management (Zong & Wu, 2013). The Computer Network Information Center of the Chinese Academy of Sciences proposed a data quality assessment method and index system (Data Application Environment Construction and Service of the Chinese Academy of Sciences, 2009) in which data quality is divided into three categories including external form quality, content quality, and the utility of quality. Each category is subdivided into quality characteristics and an evaluation index.

In summary, the existing studies focus on two aspects: a series of studies of web data quality and studies in specific areas, such as biology, medicine, geophysics, telecommunications, scientific data, etc. Big data as an emerging technology, acquires more and more attention but also lacks research results in establishing big data quality and assessment methods under multi-source, multi-modal environments (Song & Qin, 2007).

3 The Challenges of Data Quality in the Big Data Era

3.1 Features of big data

Because big data presents new features, its data quality also faces many challenges. The characteristics of big data come down to the 4Vs: Volume, Velocity, Variety, and Value (Katal, Wazid, & Goudar, 2013). Volume refers to the tremendous volume of the data. We usually use TB or above magnitudes to measure this data volume. Velocity means that data are being formed at an unprecedented speed and must be dealt with in a timely manner. Variety indicates that big data has all kinds of data types, and this diversity divides the data into structured data and unstructured data. These multityped data need higher data processing capabilities.

Finally, Value represents low-value density. Value density is inversely proportional to total data size, the greater the big data scale, the less relatively valuable the data.

3.2 The challenges of data quality

Because big data has the 4V characteristics, when enterprises use and process big data, extracting high-quality and real data from the massive, variable, and complicated data sets becomes an urgent issue. At present, big data quality faces the following challenges:

- ***The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.***

In the past, enterprises only used the data generated from their own business systems, such as sales and inventory data. But now, data collected and analyzed by enterprises have surpassed this scope. Big data sources are very wide, including: 1) data sets from the internet and mobile internet (Li & Liu, 2013); 2) data from the Internet of Things; 3) data collected by various industries; 4) scientific experimental and observational data (Demchenko, Grosso & Laat, 2013), such as high-energy physics experimental data, biological data, and space observation data. These sources produce rich data types. One data type is unstructured data, for example, documents, video, audio, etc. The second type is semi-structured data, including: software packages/modules, spreadsheets, and financial reports. The third is structured data. The quantity of unstructured data occupies more than 80% of the total amount of data in existence.

As for enterprises, obtaining big data with complex structure from different sources and effectively integrating them are a daunting task (McGilvray, 2008). There are conflicts and inconsistent or contradictory phenomena among data from different sources. In the case of small data volume, the data can be checked by a manual search or programming, even by ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform). However, these methods are useless when processing PB-level even EB-level data volume.

- ***Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.***

After the industrial revolution, the amount of information dominated by characters doubled every ten years. After 1970, the amount of information doubled every three years. Today, the global amount of information can be doubled every two years. In 2011, the amount of global data created and copied reached 1.8 ZB. It is difficult to collect, clean, integrate, and finally obtain the necessary high-quality data within a reasonable time frame. Because the proportion of unstructured data in big data is very high, it will take a lot of time to transform unstructured types into structured types and further process the data. This is a great challenge to the existing techniques of data processing quality.

- ***Data change very fast and the “timeliness” of data is very short, which necessitates higher requirements for processing technology.***

Due to the rapid changes in big data, the “timeliness” of some data is very short. If companies can't collect the required data in real time or deal with the data needs over a very long time, then they may obtain outdated and invalid information. Processing and analysis based on these data will produce useless or misleading conclusions, eventually leading to decision-making mistakes by governments or enterprises. At present, real-time processing and analysis software for big data is still in development or improvement phases; really effective commercial products are few.

- ***No unified and approved data quality standards have been formed in China and abroad, and research on the data quality of big data has just begun.***

In order to guarantee the product quality and improve benefits to enterprises, in 1987 the International Organization for Standardization (ISO) published ISO 9000 standards. Nowadays, there are more than 100 countries and regions all over the world actively carrying out these standards. This implementation promotes mutual understanding among enterprises in domestic and international trade and brings the benefit of eliminating trade barriers. By contrast, the study of data quality standards began in the 1990s, but not until 2011 did ISO published ISO 8000 data quality standards (Wang, Li, & Wang, 2010). At present, more than 20 countries have participated in this standard, but there are many disputes about it. The standards need to be mature and perfect. At the same time, research on big data quality in China and abroad has just begun and there are, as yet, few results.

4 Quality Criteria of Big Data

Big data is a new concept, and academia hasn’t made a uniform definition of its data quality and quality criteria. The literature differs on a definition of data quality, but one thing is certain: data quality depends not only on its own features but also on the business environment using the data, including business processes and business users. Only the data that conform to the relevant uses and meet requirements can be considered qualified (or good quality) data. Usually, data quality standards are developed from the perspective of data producers. In the past, data consumers were either direct or indirect data producers, which ensured the data quality. However, in the age of big data, with the diversity of data sources, data users are not necessarily data producers. Thus, it is very difficult to measure data quality. Therefore, we propose a hierarchical data quality standard from the perspective of the users, as shown in **Figure 1**.

We chose data quality dimensions commonly accepted and widely used as big data quality standards and redefined their basic concepts based on actual business needs. At the same time, each dimension was divided into many typical elements associated with it, and each element has its own corresponding quality indicators. In this way, hierarchical quality standards for big data were used for evaluation. **Figure 2** shows a universal two-layer data quality standard. Some detailed data quality indicators are given in **Table 1**.

In **Figure 2**, the data quality standard is composed of five dimensions of data quality - availability, usability, reliability, relevance, and presentation quality. For each dimension, we identified 1–5 elements with good practices. The first four quality dimensions are regarded as indispensable, inherent features of data quality, and the final dimension is additional properties that improve customer satisfaction. Availability is defined as the degree of convenience for users to obtain data and related information, which is divided into the three elements of accessibility, authorization, and timeliness. The concept of usability means whether the data are useful and meet users’ needs, including data definition/documentation, reliability, and meta-data. Reliability refers to whether we can trust the data; this consists of accuracy, consistency, completeness, and auditability.

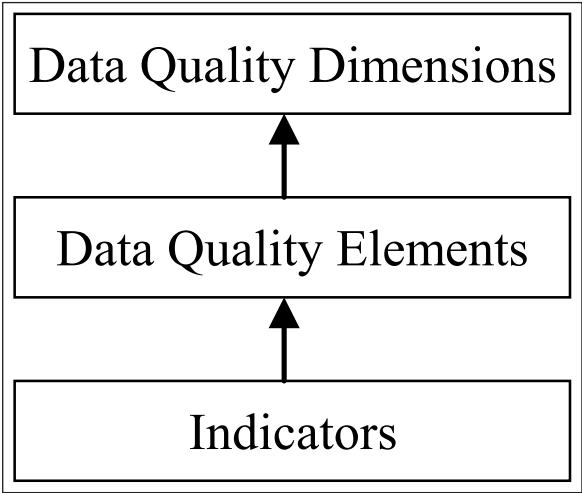


Figure 1: Data quality framework.

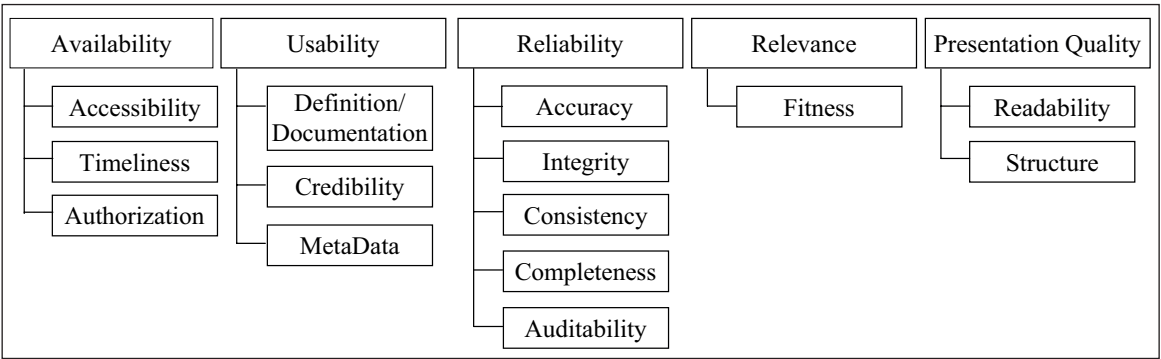


Figure 2: A universal, two-layer big data quality standard for assessment.

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	<ul style="list-style-type: none"> ■ Whether a data access interface is provided ■ Data can be easily made public or easy to purchase
	2) Timeliness	<ul style="list-style-type: none"> ■ Within a given time, whether the data arrive on time ■ Whether data are regularly updated ■ Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	<ul style="list-style-type: none"> ■ Data come from specialized organizations of a country, field, or industry ■ Experts or specialists regularly audit and check the correctness of the data content ■ Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	<ul style="list-style-type: none"> ■ Data provided are accurate ■ Data representation (or value) well reflects the true state of the source information ■ Information (data) representation will not cause ambiguity
	2) Consistency	<ul style="list-style-type: none"> ■ After data have been processed, their concepts, value domains, and formats still match as before processing ■ During a certain time, data remain consistent and verifiable ■ Data and the data from other data sources are consistent or verifiable
	3) Integrity	<ul style="list-style-type: none"> ■ Data format is clear and meets the criteria ■ Data are consistent with structural integrity ■ Data are consistent with content integrity
	4) Completeness	<ul style="list-style-type: none"> ■ Whether the deficiency of a component will impact use of the data for data with multi-components ■ Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	<ul style="list-style-type: none"> ■ The data collected do not completely match the theme, but they expound one aspect ■ Most datasets retrieved are within the retrieval theme users need ■ Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	<ul style="list-style-type: none"> ■ Data (content, format, etc.) are clear and understandable ■ It is easy to judge that the data provided meet needs ■ Data description, classification, and coding content satisfy specification and are easy to understand

Table 1: The hierarchical big data quality assessment framework (partial content).

adequacy, and auditability elements. Relevance is used to describe the degree of correlation between data content and users' expectations or demands; adaptability is its quality element (Cappiello, Francalanci, & Pernici, 2004). Presentation quality refers to a valid description method for the data, which allows users to fully understand the data. Its dimensions are readability and structure. Descriptions of the data quality elements are given below.

• **Accessibility**

Accessibility refers to the difficulty level for users to obtain data. Accessibility is closely linked with data openness, the higher the data openness degree, the more data types obtained, and the higher the degree of accessibility.

• **Timeliness**

Timeliness is defined as the time delay from data generation and acquisition to utilization (McGivray, 2010). Data should be available within this delay to allow for meaningful analysis. In the age of big data, data content changes quickly so timeliness is very important.

• **Authorization**

Authorization refers to whether an individual or organization has the right to use the data.

- **Credibility**

Credibility is used to evaluate non-numeric data. It refers to the objective and subjective components of the believability of a source or message. Credibility of data has three key factors: reliability of data sources, data normalization, and the time when the data are produced.

- **Definition/Documentation**

Definition/document consists of data specification, which includes data name, definition, ranges of valid values, standard formats, business rules, etc. Normative data definition improves the degree of data usage.

- **MetaData**

With the increase of data sources and data types, because data consumers distort the meaning of common terminology and concepts of data, using data may bring risks. Therefore, data producers need to provide metadata describing different aspects of the datasets to reduce the problems caused by misunderstanding or inconsistencies.

- **Accuracy**

To ascertain the accuracy of a given data value, it is compared to a known reference value. In some situations, accuracy can be easily measured, such as gender, which has only two definite values: male and female. But in other cases, there is no known reference value, making it difficult to measure accuracy. Because accuracy is correlated with context to some extent, data accuracy should be decided by the application situation.

- **Consistency**

Data consistency refers to whether the logical relationship between correlated data is correct and complete. In the field of databases (Silberschatz, Korth, & Sudarshan, 2006), it usually means that the same data that are located in different storage areas should be considered to be equivalent. Equivalency means that the data have equal value and the same meaning or are essentially the same. Data synchronization is the process of making data equal.

- **Integrity**

The term data integrity is broad in scope and may have widely different meanings depending on the specific context. In a database, data with “integrity” are said to have a complete structure. Data values are standardized according to a data model and/or data type. All characteristics of the data must be correct – including business rules, relations, dates, definitions, etc. In information security, data integrity means maintaining and assuring the accuracy and consistency of data over its entire life-cycle. This means that data cannot be modified in an unauthorized or undetected manner.

- **Completeness**

If a datum has multiple components, we can describe the quality with completeness. Completeness means that the values of all components of a single datum are valid. For example, for image color, RGB can be used to describe red, green, and blue, and RGB represents all parts of the color data. If the color value of a certain component is missing, the image cannot show the real color and its completeness is destroyed (Wang & Storey, 1995).

- **Auditability**

From the perspective of audit application, the data life cycle includes three phases: data generation, data collection, and data use (Wang & Zhu, 2007). But here auditability means that auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase.

- **Fitness**

Fitness has two-level requirements: 1) the amount of accessed data used by users and 2) the degree to which the data produced matches users' needs in the aspects of indicator definition, elements, classification, etc.

- **Readability**

Readability is defined as the ability of data content to be correctly explained according to known or well defined terms, attributes, units, codes, abbreviations, or other information.

- **Structure**

More than 80% of all data is unstructured, therefore, structure refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology.

We present a big data quality assessment framework in **Table 1**, which lists the common quality elements and their associated indicators. Generally, a quality element has its own multi-indicators.

5 QUALITY ASSESSMENT PROCESS FOR BIG DATA

An appropriate quality assessment method for big data is necessary to draw valid conclusions. In this paper, we propose an effective data quality assessment process with a dynamic feedback mechanism based on big data's own characteristics, shown in **Figure 3**.

Determining the goals of data collection is the first step of the whole assessment process. Big data users rationally choose the data to be used according to their strategic objectives or business requirements, such as operations, decision making, and planning. The data sources, types, volume, quality requirements, assessment criteria, and specifications as well as the expected goals need to be determined in advance.

In different business environments, the selection of data quality elements will differ. For example, for social media data, timeliness and accuracy are two important quality features. However, because it is difficult to directly judge accuracy (Shankaranarayanan, Ziad, & Wang, 2012), some additional information is needed to judge the raw data, and other data sources serve as supplements or evidence. Therefore, credibility has become an important quality dimension. However, social media data are usually unstructured, and their consistency and integrity are not suitable for evaluation. The field of biology is an important source of big data. However, due to the lack of uniform standards, data storage software and data formats vary widely. Thus, it is difficult to regard consistency as a quality dimension, and the needs of regarding timeliness and completeness as data quality dimensions are not high.

In order to further quality assessment, we need to choose specific assessment indicators for every dimension. These require the data to comply with specific conditions or features. The formulation of assessment indicators also depends on the actual business environment.

Each quality dimension needs different measurement tools, techniques, and processes, which leads to differences in assessment times, costs, and human resources. In a clear understanding of the work required to assess each dimension, choosing those dimensions that meet the needs can well define a project's scope. The preliminary assessment results of data quality dimensions determine the baseline while the remaining assessment as a part of the business process is used for continuous detection and information improvement.

After the quality assessment preparation is completed, the process enters the data acquisition phase. There are many ways to collect data (Zhu & Xiong, 2009), including: data integration, search-download, web

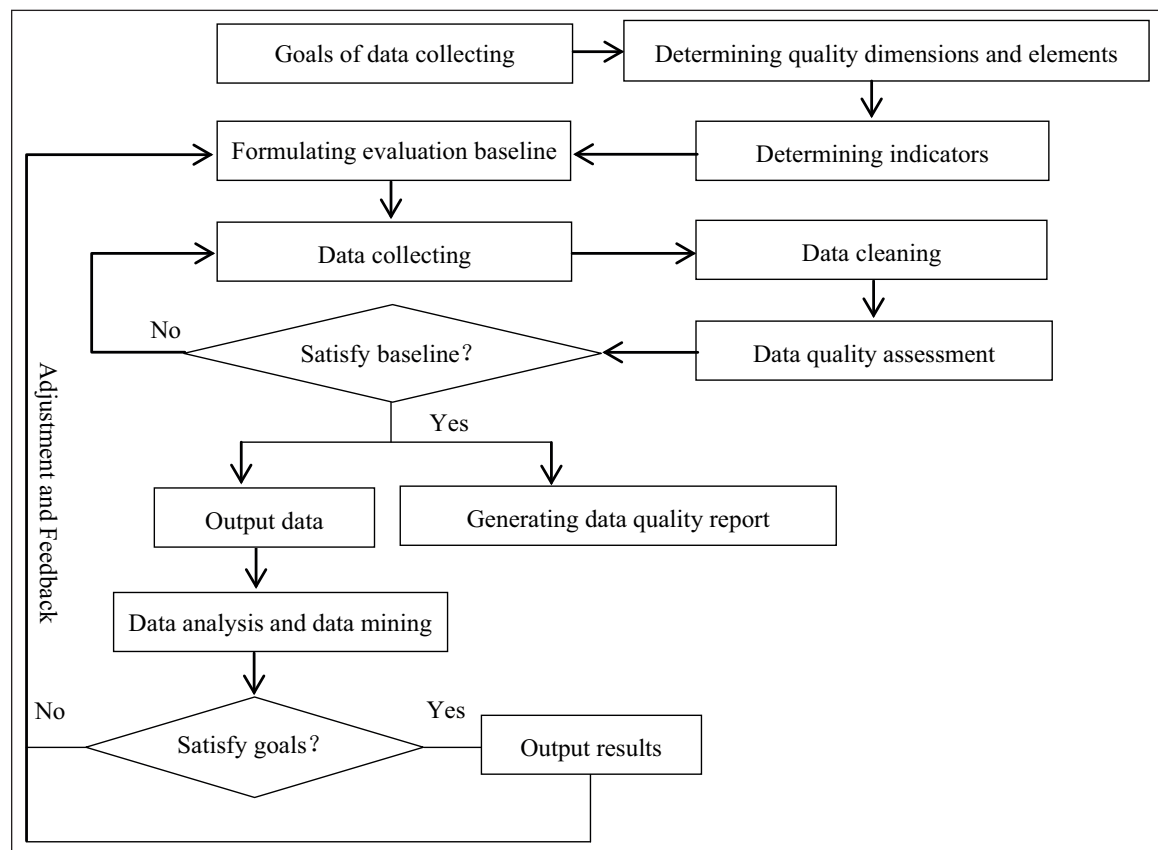


Figure 3: Quality assessment process for big data.

crawlers, agent methods, carrier monitors, etc. In the age of big data, data acquisition is relatively easy, but much of the data collected is not always good. We need to improve data quality as far as possible under these conditions without a large increase in acquisition cost.

Big data sources are very wide and data structures are complex. The data received may have quality problems, such as data errors, missing information, inconsistencies, noise, etc. The purpose of data cleaning (data scrubbing) is to detect and remove errors and inconsistencies from data in order to improve their quality. Data cleaning can be divided into four patterns based on implementation methods and scopes (Wang, Zhang, & Zhang, 2007): manual implementation, writing of special application programs, data cleaning unrelated to specific application fields, and solving the problem of a type of specific application domain. In these four approaches, the third has good practical value and can be applied successfully.

Then, the process enters the data quality assessment and monitoring phases. The core of data quality assessment is how to evaluate each dimension. The current method has two categories: qualitative and quantitative methods. The qualitative evaluation method is based on certain evaluation criteria and requirements, according to assessment purposes and user demands, from the perspective of qualitative analysis to describe and assess data resources. Qualitative analysis should be performed by subject experts or professionals. The quantitative method is a formal, objective, and systematic process in which numerical data are utilized to obtain information. Therefore, objectivity, generalizability, and numbers are features often associated with this method, whose evaluation results are more intuitive and concrete.

After assessment, the data can be compared with the baseline for the data quality assessment established above. If the data quality accords with the baseline standard, a follow-up data analysis phase can be entered, and a data quality report will be generated. Otherwise, if the data quality fails to satisfy the baseline standard, it is necessary to acquire new data.

Strictly speaking, data analysis and data mining do not belong to the scope of big data quality assessment, but they play an important role in the dynamic adjustment and feedback of data quality assessment. We can use these two methods to discover whether valuable information or knowledge exists in big data and whether the knowledge can be helpful for policy proposals, business decisions, scientific discoveries, disease treatments, etc. If the analysis results meet the goal, then the results are outputted and fed back to the quality assessment system so as to provide better support for the next round of assessment. If results do not reach the goal, the data quality assessment baseline may not be reasonable, and we need to adjust it in a timely fashion in order to obtain results in line with our goals.

6 Conclusion

The arrival of the big data era makes data in various industries and fields present explosive growth. How to ensure big data quality and how to analyze and mine information and knowledge hidden behind the data become major issues for industry and academia. Poor data quality will lead to low data utilization efficiency and even bring serious decision-making mistakes. We analyzed the challenges faced by big data quality and proposed the establishment and hierarchical structure of a data quality framework. Then, we formulated a dynamic big data quality assessment process with a feedback mechanism, which has laid a good foundation for further study of the assessment model. The next stage of research will involve the construction of a big data quality assessment model and formation of a weight coefficient for each assessment indicator. At the same time, the research team will develop an algorithm used to make a practical assessment of the big data quality in a specific field.

7 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61170096, the Major Program of National Natural Science Foundation of China under Grant No. 71331005, the Shanghai Science and Technology Development Funds under Grant No. 13dz2260200, 13511504300, and the Department of Science and Technology of Yunnan Province under Grant No. 2012FD004.

8 References


- Alan, F. K., Sanil, A. P., Sacks, J., et al. (2001) Workshop Report: Affiliates Workshop on Data Quality, North Carolina: NISS.
- Alexander, J. E., & Tate, M. A. *Web wisdom: How to evaluate and create information on the web*, Mahwah, NJ: Erlbaum.
- Cao, J. J., Diao, X. C., Wang, T., et al. (2010) Research on Some Basic Problems in Data Quality Control. *Micro-computer Information* 09, pp 12–14.
- Cappiello, C., Francalanci, C., & Pernici, B. (2004) Data quality assessment from user's perspective. *Procedures of the 2004 International Workshop on Information Quality in Information Systems*, New York: ACM, pp 78–73.
- Crosby, P. B. (1988) *Quality is Free: The Art of Making Quality Certain*, New York: McGraw-Hill.
- Data Application Environment Construction and Service of Chinese Academy of Sciences (2009) Data Quality Evaluation Method and Index System. Retrieved October 30, 2013 from the World Wide Web: <http://www.csdb.cn/upload/101205/1012052021536150.pdf>
- Demchenko, Y., Grosso, P., de Laat, C., et al. (2013) Addressing Big Data Issues in Scientific Data Infrastructure. *Procedures of the 2013 International Conference on Collaboration Technologies and Systems*, California: ACM, pp 48–55.
- Feng, Z. Y., Guo, X. H., Zeng, D. J., et al. (2013) On the research frontiers of business management in the context of Big Data. *Journal of Management Sciences in China* 16(01), pp 1–9.
- Gantz, J., & Reinsel, D. (2012) THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Retrieved February, 2013 from the World Wide Web: <http://www.emc.com/collateral/analyst-reports/idc-digital-universe-western-europe.pdf>
- General Administration of Quality Supervision (2008) Inspection and Quarantine of the People's Republic of China. *Quality management systems-Fundamentals and vocabulary* (GB/T19000–2008/ISO9000:2005), Beijing.
- Katal, A., Wazid, M., & Goudar, R. (2013) Big Data: Issues, Challenges, Tools and Good Practices. *Procedures of the 2013 Sixth International Conference on Contemporary Computing*, Noida: IEEE, pp 404–409.
- Katerattanakul, P., & Siau, K. (1999) Measuring information quality of web sites: Development of an instrument. *Procedures of the 20th International Conference on Information Systems*, North Carolina: ACM, pp 279–285.
- Knight, S., & Burn, J. (2005) Developing a Framework for Assessing Information Quality on the World Wide Web. *Information Science Journal* 18, pp 159–171.
- Li, G. J., & Chen, X. Q. (2012) Research Status and Scientific Thinking of Big Data. *Bulletin of Chinese Academy of Sciences* 27(06), pp 648–657.
- Li, J. Z., & Liu, X. M. (2013) An Important Aspect of Big Data: Data Usability. *Journal of Computer Research and Development* 50(6), pp 1147–1162.
- McGilvray, D. (2008) *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*, California: Morgan Kaufmann.
- McGilvray, D. (2010) *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*, Beijing: Publishing House of Electronics Industry.
- Meng, X. F., & Ci, X. (2013) Big Data Management: Concepts, Techniques and Challenges. *Journal of Computer Research and Development* 50(1), pp 146–169.
- Nature (2008) Big Data. Retrieved November 5, 2013 from the World Wide Web: <http://www.nature.com/news/specials/bigdata/index.html>
- Science (2011) Special online collection: Dealing with data. Retrieved November 5, 2013 from the World Wide Web: <http://www.sciencemag.org/site/special/data/>
- Shankaranarayanan, G., Ziad, M., & Wang, R. Y. (2012) Preliminary Study on Data Quality Assessment for Socialized Media. *China Science and Technology Resources* 44(2), pp 72–79.
- Shanks, G., & Corbitt, B. (1999) Understanding data quality: Social and cultural aspects. *Procedures of the 10th Australasian Conference on Information Systems*, Wellington: MCB University Press Ltd., pp 785–797.
- Silberschatz, A., Korth, H., & Sudarshan, S. (2006) *Database System Concepts*, Beijing: Higher Education Press.
- Song, M., & Qin, Z. (2007) Reviews of Foreign Studies on Data Quality Management. *Journal of Information* 2, pp 7–9.
- Wang, H., & Zhu, W. M. (2007) Quality of Audit Data: A Perspective of Evidence. *Journal of Nanjing University (Natural Sciences)* 43(1), pp 29–34.

- Wang, J. L., Li, H., & Wang, Q. (2010) Research on ISO 8000 Series Standards for Data Quality. *Standard Science* 12, pp 44–46.
- Wang, R., & Storey, V. (1995) Framework for Analysis of Quality Research. *IEEE Transactions on Knowledge and Data Engineering* 1(4), pp 623–637.
- Wang, R. Y., & Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12(4), pp 5–33.
- Wang, Y. F., Zhang, C. Z., Zhang, B. B., et al. (2007) A Survey of Data Cleaning. *New Technology of Library and Information Service* 12, pp 50–56.
- Zhu, X., & Gauch, S. (2000) Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. *Procedures of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens: ACM, pp 288–295.
- Zhu, Y. Y., & Xiong, Y. (2009) *Datalogy and Data Science*, Shanghai: Fudan University Press.
- Zong, W., & Wu, F. (2013) The Challenge of Data Quality in the Big Data Age. *Journal of Xi'an Jiaotong University (Social Sciences)* 33(5), pp 38–43.

How to cite this article: Cai, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14: 2, pp.1-10, DOI: <http://dx.doi.org/10.5334/dsj-2015-002>

Published: 22 May 2015

Copyright: © 2015 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License (CC-BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/3.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 